




## Reconstructing historical texts from fragmentary sources: Charles S. Parnell and the Irish crisis, 1880-86

Eugenio Biagini <efb21\_at\_cam\_dot\_ac\_dot\_uk>, University of Cambridge  <https://orcid.org/0000-0002-7321-0434?lang=en>

Patrick Geoghegan <GEOGHANP\_at\_tcd\_dot\_ie>, Trinity College Dublin  <https://orcid.org/0000-0002-0605-8604?lang=en>

Hugh Hanley <hdh27\_at\_cam\_dot\_ac\_dot\_uk>, University of Cambridge  <https://orcid.org/0000-0003-0209-0081>

Aneirin Jones <aneirinjones\_at\_hotmail\_dot\_com>, University of Cambridge

Huw Jones <hej23\_at\_cam\_dot\_ac\_dot\_uk>, University of Cambridge  <https://orcid.org/0000-0002-8533-9083>

### Abstract

Charles Stewart Parnell was one of the most controversial and effective leaders in the United Kingdom in the second half of the nineteenth century. Almost single-handedly, he transformed the proposal of Home Rule for Ireland from a languishing irrelevance to a mass-supported cause. Though the historiography on Parnell is substantial, his speeches – the main primary sources for accessing both his thinking and strategies – have never been collected or edited. One of the core questions in working towards an edition of his speeches was whether it would be possible to use automated methods on these fragmentary sources to reconstruct what Parnell actually said in them. We were also interested in how the reports varied, and what that variation might tell us about the practices and biases of the journalists who wrote them and the newspapers which published them. This article discusses the use of two digital tools in our attempts to answer these research questions: CollateX, which was designed by Digital Humanities practitioners for the comparison of textual variants, and SBERT Sentence Transformers, which establishes levels of similarity between texts. In this article we talk about how the application of digital methods to the corpus led us away from the idea of producing definitive reconstructions of the speeches, and towards a deeper understanding of the corpus and the journalistic practices which went into its creation.

## 1 Introduction

Charles Stewart Parnell (1846-1891) was one of the most controversial and effective leaders in the United Kingdom in the second half of the nineteenth century. Almost single-handedly, he transformed the proposal of Home Rule for Ireland from a languishing irrelevance to a mass-supported cause. The political backing which he secured was sufficient to persuade one of the two major British parties, W.E. Gladstone's Liberals, to adopt Home Rule from 1886. Their opponents, the Tories, felt compelled to respond to the extent that they eventually redefined their identity in reaction to the Parnell programme, and became the Conservative and Unionist party, and yet were forced to accept and implement some of Parnell's demands, specifically with reference to land reform and the democratisation of local government. What is even more extraordinary is that the remapping of the United Kingdom's political landscape that Parnell provoked proved long-lasting, if not permanent. The idea of Home Rule survived both Parnell himself and Gladstone and came to epitomise the quintessential British and Irish road to constitutional reform from 1920 (when it was adopted for Northern Ireland) and 1998 (when it was applied, in a modified form, to both Scotland and Wales) [Jackson 2003] [Jackson 2012]. Moreover, Parnell's parallel campaign for land reform not only succeeded beyond his own hopes and expectations, but was also exported to other parts of the British Empire, providing the blueprint for the negotiation of late-colonial agrarian conflicts from India to Kenya and Uganda [Low 1991].

Though the historiography on Parnell is substantial (e.g. [Boyce and O'Day 1991]; [Travers and McCartney 2013]), his speeches – the main primary sources for accessing both his thinking and strategies – have never been collected or edited. Though they have been frequently cited, they have not been systematically studied and have often been quoted selectively from newspaper reports which have been considered to be at best biased and sometimes tendentious. In 2020 we secured a Cambridge Humanities Research Grant to remedy this situation by working towards the first critical edition of Parnell's speeches.

If Parnell kept drafts of his speeches we do not have them, and our sources consist of multiple reports made by the newspapers that recorded his words and commented on them. One of the core questions in working towards an edition was whether it would be possible to use automated methods on these fragmentary sources to reconstruct what Parnell actually said in his speeches. We were also interested in how the reports varied, and what that variation might tell us about the practices and biases of the journalists who wrote them and the newspapers which published them. This article discusses the use of two digital tools in our attempts to answer these research questions: CollateX, which was designed by digital humanities practitioners for the comparison of textual variants, and SBERT Sentence Transformers, which establishes levels of similarity between texts.

Perhaps the most interesting aspect of the project was that the digital methods did not generally reveal the implicit bias in reporting the content of the speeches that we were looking for (with one significant exception, described in Section 6.2), nor did they allow us to construct single reliable texts of the speeches. Instead, they gave us insight into the process of reporting Parnell's speeches, and how the practices and methods of newspapers, editors and journalists, their approaches, mistakes and omissions, might help historians to better understand the speeches as a corpus. This highlights the role of digital humanities methods and approaches in generating new perspectives, even if they appear to be unsuccessful in answering our original questions – as Willard McCarty has said, “a good model can be fruitful in two ways: either by fulfilling our expectations, and so strengthening its theoretical basis, or by violating them, and so bringing that basis into question. [...] from the research perspective [...] failure to give us what we expect is by far the most important result” [McCarty 2013].

## 2 Related Work

Related work on automated approaches to digital editions highlights two aspects of particular relevance to our project: the edition as a process rather than a finished output, and the necessity of combining digital methods with the kind of detailed interpretation and deep reading usually associated with more traditional humanities research.

The adaptation of the scholarly edition to the digital age has led to an emphasis on viewing the text within the context of its creation, and an understanding of textual boundaries as both mutable and extensible. Recent approaches to digital scholarly editing position editorial work as a dynamic, creative process, with hybridity and differing textual versions given a new prominence [Nabugodi and Ohge 2022]. This is also true of the edition in itself, which is increasingly approached as a multifaceted object, with a “final” text presented alongside manuscript variants, related texts, images and digital tools. This movement towards a mutable vision of the text has a clear relevance to what might be seen as the “limitations” of the Parnell speech source material – the ephemeral nature of the speech itself, and the lack of any canonical version of the text against which to compare witnesses. While recent literature emphasises the importance of digital approaches to the scholarly edition, it is also clear that editorial intervention is an important factor, both in the interpretation of results and in feeding back to the digital processes themselves.

The blurring of the boundary between an authoritative central text and its context is a problem which increasingly preoccupies the producers of modern scholarly editions. James Cummings draws a distinction between the “document” as “a particular instance of a physical manifestation of this text” and the “work” as “an abstraction as understood by readers (including authors and editors)” [Cummings 2019a]. From this viewpoint, the edition is continually destabilised by the contextual environment from which it originates and continues to be formed. While in some ways this is true of all editions, it is presented as being particularly true of the digital edition, which has the potential “[to] be near-infinitely refactorable and dynamically to provide different views depending on external interactions” [Cummings 2019a]. This can be expanded to the process of creating the digital edition and the foregrounding within the edition itself of the research methods and techniques, such as collation, which are used as part of the editorial process. Dirk van Hulle asserts the potential of modern editions to simulate “a process, such as the creative and imaginative process of a literary work” [Van Hulle 2019]. This reflects our own experience of the automated collation of an

edition as a research method which raises new questions about the corpus, rather than a tool for providing definitive versions of texts.

CollateX is the most commonly used software for the collation of texts in the digital humanities. It was conceived within the Interedition research group, a cross-institutional initiative created with the aim of developing tools for textual scholarship in a collaborative environment. In "Computer-supported collation of modern manuscripts: CollateX and the Beckett Digital Manuscript Project," members of the group outline the implementation of the software in relation to a digital edition of Samuel Beckett's manuscripts [Dekker et al. 2014]. Their approach is based on the "Gothenburg Model", created to explore the "conceptual commonalities" between fields relating to collation in digital textual scholarship. Here, the collation process is broken up into five steps:

- Tokenisation of the texts to be compared into textual sub-units such as characters words or sentences
- Normalisation of the tokens to ensure that "equivalent" tokens will align correctly (by contrast, see [Birnbbaum and Spadini 2020] on normalisation as a process that occurs at every stage of collation, including transcription of witnesses)
- Alignment of tokens between texts to see where they match and differ
- Analysis of the computed alignment to interpret and correct it
- Output/visualisation of the collation results

Much emphasis is put on the human aspect of this process, both in the analysis of results and in decisions on the appropriate level of tokenisation. The inherent ambiguity of the collation process is also highlighted, particularly in relation to transposed text: "In some cases, even human interpretation may of course not determine decisively whether an actual transposition took place. We may have to conclude that some cases of potential transposition cannot be determined with absolute certainty" [Dekker et al. 2014]. This reciprocal and iterative relationship between digital methods and scholarly interpretation, including establishing the point at which the methods fail to produce conclusive results, was central to our work.

For our project, in which the base text of the speech (i.e. what Parnell actually said) is absent, perhaps the most relevant examples of related research come from work on medieval manuscripts and biblical editions, where multiple sources are collated in an ongoing effort to reconstruct an authoritative text. The *Novum Testamentum Graecum: Editio Critica Maior* is an ambitious project that "has as its goal to offer a new reconstruction of the earliest attainable text for each of the New Testament writings, termed the *Ausgangstext* or *Initial Text*, and to present the evidence for the textual history of the Greek New Testament during the first millennium" [Houghton et al. 2020]. The group also use CollateX, taking advantage of its concept of "a baseless collation, allowing the divergences in the textual tradition to be presented without assumptions about the earliest form of text" [Houghton et al. 2020]. While automated collation and digital tools have a prominent role in the project, these are combined with editorial procedures that emphasise the human, interpretative aspects of textual scholarship. Automated outputs are rigorously checked for misalignment and "spelling differences, errors or other peculiarities of individual manuscripts which are considered to be 'noise' and are not deemed to be significant for the edition" are eliminated [Houghton et al. 2020]. It is made clear that any deployment of digital tools in relation to an edition, particularly in the absence of a base text, must form part of a collaborative editorial workflow.

### 3 Dataset

Our dataset consists of 630 TEI P5 XML records relating to reports of Parnell's speeches. These were created by members of the project team through a combination of OCR and manual transcription. For each report, metadata is recorded on the newspaper (or other outlet) where it was published, on the date and place of publication, and on the type of report (newspaper, pamphlet etc.). Identifier schemes are used for places (Getty Thesaurus of Geographical Names), and publications (Virtual International Authority File, or VIAF) to disambiguate entries and to enable future linked data approaches. ISO 8601 forms of dates are recorded alongside transcriptions of the dates as they appear in the reports. While the hierarchical nature of TEI has come under some criticism in the context of literary editions (e.g. [McGann 2022], but see rebuttal of some points in [Cummings 2019b]), its structured approach has lent itself very well to the aims of our project, in particular the close relationship between detailed metadata and transcription. TEI also provides the kind of general standardisation which could allow for comparative approaches with related datasets.

A separate local authority file is maintained for the speeches themselves, recording when and where each speech took place (if known) and a short summary of the content of the speech (where applicable). Each speech is given an identifier which is referenced in the report records - allowing for the grouping together of all reports relating to a single speech, and creating an entry point where new information about speeches (as distinct from reports of speeches) can be recorded.

Our TEI records also contain the texts of the reports. Other than basic structural units such as headings and paragraphs we have not introduced further markup (e.g. of places, people, dates) into the report texts. This provides a blank slate for textual analysis and opens up the possibility of using natural language processing techniques such as parts-of-speech tagging and named entity recognition in future stages of the project.

The data has been modelled both for publication and research - allowing us to easily extract reports relating to speeches, and to analyse and compare them by date, place and publication. This involves, for instance, comparing reports published in English newspapers with those published in Irish newspapers, looking at aspects of the speeches over time, seeing how Parnell tailored his speech to different audiences or places.

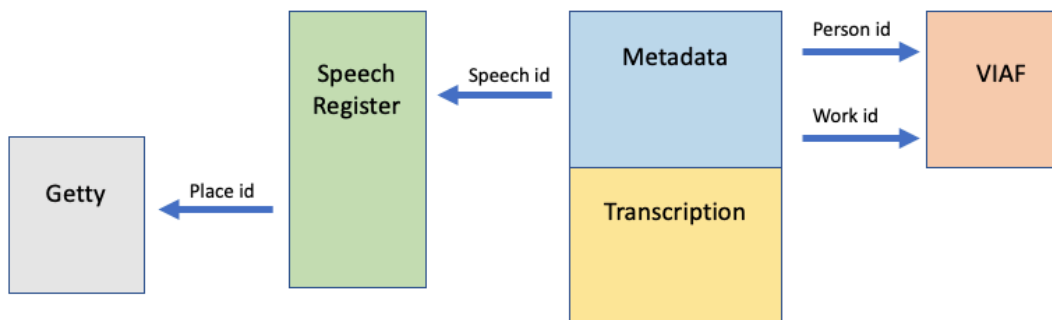


Figure 1. TEI data model with links to external authority schemes

### 4 Methodology

A key problem in the development of our methodology was finding texts with sufficient levels of similarity for the collation process to be effective. Reports for the same speech sometimes differed to the point where any kind of automated textual comparison became impossible, especially in the comparison of full transcriptions of speeches with summaries of their content. To address this problem, an additional step was introduced before the collation process to identify which texts would collate effectively. This initial stage produced interesting results on the general level of similarity between accounts of speeches, pointing to patterns of copying and adaptation in the writing of reports, and also absence of reporting

or partial coverage by some newspapers. This process of calculating source similarity, which was first seen as a purely pragmatic activity to assemble reports which were suitable for collation, instead opened up interesting new pathways for research on journalistic practice around reporting on Parnell.

Reports which met the required level of similarity for collation were processed using CollateX. The outputs of the collation process were assessed by subject experts to see how effective the workflow was in providing us with useful insights not only into what Parnell did or did not say, but also on the way in which newspapers and other outlets reported on his speeches. One unintended result of the collation process, discussed below, was to discover that it was useful in highlighting what seemed to be the results of mishearings or misunderstandings in the contemporary transcriptions produced by journalists who were witnesses to the speeches.

15

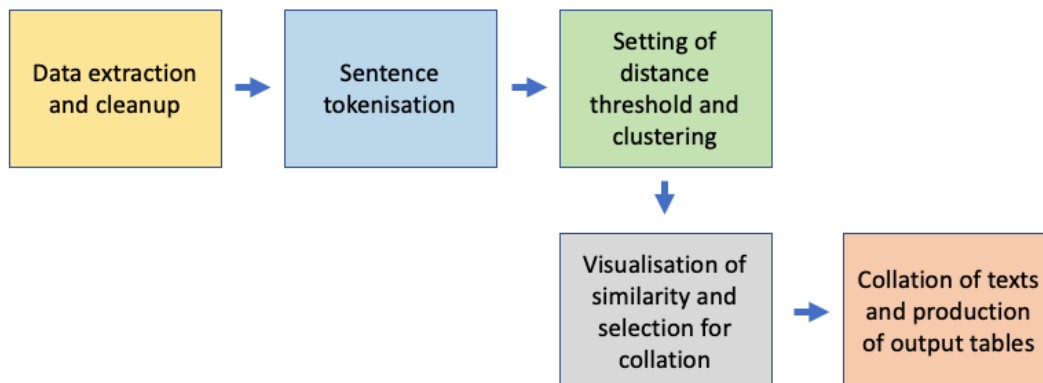


Figure 2. Project workflow

## 4.1 Selection of tools and Python libraries

Our approach was developed around the use of two Python libraries: SBERT Sentence Transformers for establishing similarity levels between sources, and CollateX for visualising the similarities and differences between them.

16

SBERT Sentence Transformers Library is a modification of the BERT (Bidirectional Encoder Representations from Transformers) model, adapting BERT to establish "semantically meaningful sentence embeddings". Sentences from a group of texts are converted into embeddings, and the cosine-similarity of these embeddings can then be used to calculate the similarity between sentences within our groups of texts.

17

CollateX is a tool specifically developed for tasks such as manuscript criticism and textual analysis. It provides a means of tokenising and comparing multiple text items, identifying similarities and differences, and aligning them in tabular output. This output format enables users to view and interpret patterns of similarity and difference between texts as they appear side by side.

18

## 4.2 Data Extraction and Sentence Tokenisation

As an initial process, we extracted the data for each speech from the speech register file and the corresponding source files related to each speech. We then used data cleaning operations to improve sentence recognition and standardise the texts – removing extra spacing, trailing spaces, newlines and preventing abbreviations (e.g. "Mr.", "Rev.", "Dr.") from ending sentences incorrectly.

19

The text for each source file was then tokenised into sentences, using a sentence tokenising tool from the NLTK Python library. For each speech we produced an intermediary dataset containing speech identification number, source, source periodical, sentence number in document and sentence text.

20

## 4.3 Sentence Transformation and Clustering

These sentences were converted into embeddings which were used to ascertain similarity levels using the SBERT Sentence Transformers tool. They were then assigned to clusters using the agglomerative clustering tool from the sklearn Python library, an algorithm which works recursively to create a hierarchical cluster tree or dendrogram of similar sentences according to a distance threshold.

21

Establishing the distance threshold was a key part of our workflow. For the purposes of our project we were aiming to capture sentences which related to the same part of the speech, requiring a high level of similarity. At a lower level of similarity (i.e. higher up the cluster tree) we were more likely to capture sentences which were similar in terms of content but did not relate to the same part of the speech.

22

Having run the code using different distance thresholds and performed checks on the results, we decided that the most appropriate threshold for our purposes was a level of 0.8. This threshold level took sentences from each report that were the same or clearly referred to the same part of the speech and gathered them together into clusters for the next stage of our process.

23

## 4.4 Speech Source Similarity

The sentence clusters we created for each speech were then used to establish the similarity of the sources relating to that speech. We did this by extracting the sentence cluster values for each source and comparing the values to those of every other source related to the same speech, creating a Jaccard matrix with calculations of similarity between the clusters for sources.

24

These matrices were then converted into a series of heatmaps for each speech, which provided a graphical representation of similarity and clusters of similarity across all the sources related to that speech.

25

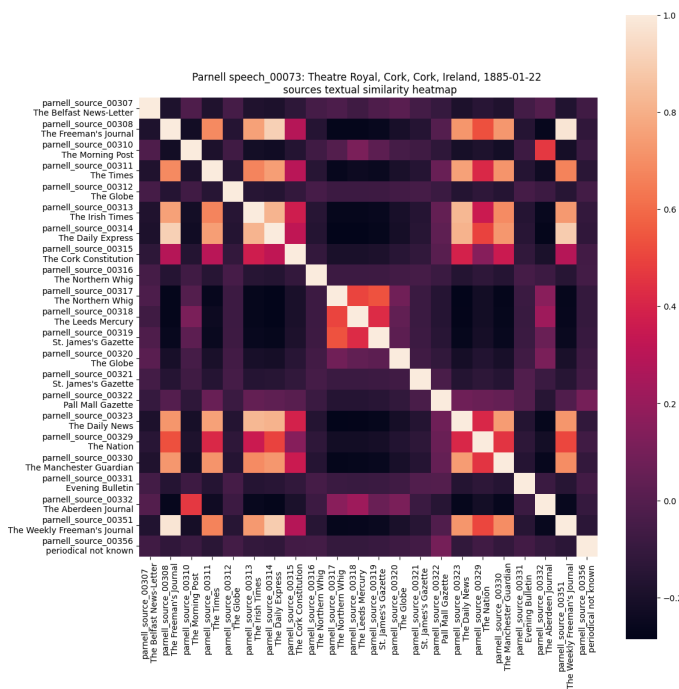


Figure 3. Example heatmap for similarity of reports of a single speech

As can be seen in Figure 3, the x and y axes have the same data and each square on the heatmap represents the level of similarity between the sentence clusters of a source on the y axis and the sentence clusters of a source on the x axis, with higher levels of similarity represented by lighter squares. The long diagonal line of lightly coloured squares represents sources which are being compared with themselves and therefore have 100 percent similarity to one another.

26

This similarity measure was not meant to be completely accurate, but rather to quickly ascertain sources with a degree of similarity which indicated that they would be good candidates for the collation process.

27

### 4.5 Collation

The input files for the collation process were selected using information gained from establishing source similarity. Once the inputs were finalised, the texts were extracted from the relevant files and normalised to remove line breaks and extra spaces.

28

The CollateX tool was then used to tokenize the text into word and punctuation tokens before aligning tokens by similarity and dissimilarity. CollateX outputs colour-coded tables of input texts, with matching rows of text given a lighter colour than non-matching rows. In order to be able to save our tables in an easily viewable form, we adapted the CollateX output using the Plotly Python library.

29

## 5 Case Study: Speech 73, Theatre Royal, Cork, 22 January 1885

This case study illustrates the importance of establishing an appropriate level of similarity between texts before embarking on automated collation.

30

### 5.1 Low Similarity Sources

Even for sources with a low level of similarity, the model seemed to be able to pick out the points where there were matches on a specific unit of text.

31

However, a coherent collation was not really possible and the CollateX model often picked out false positives using common words. For instance, isolated instances of punctuation or words such as “the”, “of” or “and” as the points of similarity between the sources.

32

parnell_source_00307 <i>The Belfast News-Letter</i>	parnell_source_00308 <i>The Freeman's Journal</i>
A numb of resolutions pledging the meeting to support Mr. Parnell	Mr. Parnell, M.P., then rose, and received a great ovation
, the	, the
Irish parliamentary party,	vast assemblage rising
and	and
the National League having been passed,	cheering for several minutes.
Mr.	Mr.
Parnell said he	Mayor and ladies and gentlemen, the mayor has kindly
claimed	claimed
their	for me your
indulgence,	indulgence,
as	and indeed last night when I set out upon the journey which
he	he
was not very well. He had been afraid that he would not be able	has described to you I felt a sinking at my heart lest when I should reach Dublin I should find myself unable to go any further, or

Table 1. Collatex output for a speech with a low level of similarity between reports

## 5.2 High Similarity Sources

Once a degree of similarity had been established, CollateX was good at picking out the similarities and differences between texts in a more coherent way, with fewer false positives. This occurred at between 40 to 60 percent sentence cluster similarity.

33

Even with sources exhibiting a very high level of cluster similarity, where there could seem to be little point in performing the collation process (as they were likely to be essentially the same), minor variations sometimes proved to be of interest, as discussed below.

34

parnell_source_00311 <i>The Times</i>	parnell_source_00315 <i>The Cork Constitution</i>
The electors who will be swamped [laughter] in the great mass of Irishmen now admitted to the rights of the constitution, so far as they exist in this country, were on the whole faithful to their trust; indeed it	It
was not until we showed by a	was not until we showed by a
good	great
many proofs	many proofs
	,
that we could do without	that we could do without
an	the
enlargement of the franchise, and	enlargement of the franchise, and
	that
with the old restricted	with the old restricted
suffrage	franchise

Table 2. Collatex output for a speech with a high level of similarity between reports

## 5.3 Working with a Cluster of Multiple Similar Sources

Looking at the heatmap for Speech 73, a main group of consistently similar sources was clearly visible: 308, 311, 313, 314, 315, 323, 330, 351. By performing a collation on these all together we could see that the model performed well in identifying similar and dissimilar passages of text.

35

However, the colour coding scheme for identifying matching and non-matching groups of sources could not be relied upon in this instance. If there were one or more non-matching pieces of text in a row that generally matched, the row was defined as non-matching and all table cells came out the same colour.

36

parnell_source_00308 <i>The Freeman's Journal</i>	parnell_source_00311 <i>The Times</i>	parnell_source_00313 <i>The Irish Times</i>	parnell_source_00314 <i>The Daily Express</i>	parnell_source_00315 <i>The Cork Constitution</i>	parnell_source_00330 <i>The Manchester Guardian</i>	parnell_source_00351 <i>The Weekly Freeman's Journal</i>
	But	But	But	as	Mr. Parnell was accompanied by	
when	when	when	when			when
I approached Ireland I found myself getting better and better	I approached Ireland I found myself getting better and better	I approached Ireland I found myself getting better and better	I approached Ireland I found myself getting better and better	I approached Ireland I found myself getting better and better		I approached Ireland I found myself getting better and better
(	(	--((	(	(		(
cheers	cheers	cheers	cheers	cheers		cheers
	),		,	),		
and cries of "		and cries of "	and cries of "			and cries of "
bravo		Bravo	bravo			bravo

Table 3. Collatex output for a speech with multiple reports

## 6 Assessment

### 6.1 Coverage and Bias

Late-Victorian polemics highlighted by modern historiography [Bew 2012] and the analysis of high-profile examples of contrasting accounts of speeches ([Travers 2000/2001], and see our own example in Section 6.2) have created the impression that the reporting of the speeches was biased and sometimes tendentious, reflecting the wish of editors and reporters to please their readers or represent Parnell in a way that would be either favourable or hostile to a certain interpretation of his words (e.g. that he was more or less constitutional or revolutionary in the way he wished to proceed with the implementation of the nationalist programme).

37

However, the results of the collation process show that the newspapers which attempted to provide full accounts of his speeches tended to agree with one another about what he had actually said, and sometimes relied on the same source (suggesting that different newspapers employed a limited number of reporters specialising on Parnell). There were discrepancies, but they generally reflected editorial decisions in cases where certain phrases were omitted or contracted for the sake of space, or where a reporter had misheard a specific word resulting in equally plausible variants, such as when *The Times* reported Parnell as saying the Conservative Party was "most remarkable for its wisdom", even though all other outlets reported it as "most remarkable for its discipline". Moreover, and not surprisingly, for the speeches that Parnell delivered when campaigning in the United States, American newspapers often provided a fuller record than their British and Irish counterparts.

38

Though the police deployed their own reporters, in general their surviving accounts relied on the records published in the newspapers, and the officers who produced or received the reports limited themselves to underlining sentences which in their view were more significant or revealing. Therefore, police records implicitly and indirectly confirmed that the newspaper press was substantially accurate and reliable in its coverage of Parnell's speeches.

39

Political bias was more clearly evident in choices on how to report (or not report) the speeches. In the results from the initial similarity analysis, it is clear that some newspapers refused to report in detail what Parnell said, providing instead short summaries. This was typically the case with Ulster Unionist newspapers. For example, *The Northern Whig*, a Belfast-based Liberal unionist publication, only published brief descriptive reports of even Parnell's most significant orations. In the five speeches in the corpus that *The Northern Whig* and *Hansard* both reported, the similarity rating given to them was zero, meaning there was no similarity between their reports. Likewise, *The Northern Whig* and the nationalist weekly, *The Irishman*, also received a score of zero over eight reports. In the reports it had in common with the nationalist daily *The Freeman's Journal* and the *Times* of London, which aspired to be perceived as the ultimate record, *The Northern Whig* had a similarity rating of 0.61 and 0.87 respectively.

40

Therefore, in our preliminary assessment of what the results tell us about how the press responded to Parnell, two considerations stand out: on the one hand, the nineteenth-century positivist emphasis on “factual” accounts [Matthew 1987] remained pervasive even when the speaker was as ambiguous and divisive as Parnell, with editors generally relegating the expression of opinion to leading articles. On the other hand, it also showed that readers relying only on regional newspapers in strongly anti-Parnell areas would not have had access to what the nationalist leader actually said, and may have tended to form their views through strongly opinionated editorials.

## 6.2 The Collation Process

The outputs from CollateX encouraged us to read the source material with a greater level of reflexivity than we had previously done. While the tool did not generally uncover obvious ideologically motivated editorial interventions, the sheer number of textual discrepancies it highlighted should persuade historians to be more circumspect about the reliability of contemporary reportage regarding what historical actors actually said.

For instance, in January 1880, during his tour of North America, the House of Representatives invited Parnell to address a House session and on the evening of 2 February a speech was given from the Speaker’s rostrum. While many media outlets covered the event, only three sources purported to give a full transcript of the speech, namely the *Congressional Record*, *The Washington Post*, and *The Irish World*.<sup>[1]</sup> However, CollateX revealed major differences between the three sources in length, wording, punctuation, and the recording of audience reactions. In terms of length, the *Congressional Record*’s report was over ten percent longer than that of *The Washington Post* and 25 percent longer than *The Irish World*. The report printed in *The Post* did not record Parnell’s formal opening in which he marked out the “[s]peaker and gentlemen of the House of Representatives” as the “ratified audience” for his remarks, and it omitted a further bulky passage where Parnell outlined the social and political context of his speech. *The Irish World* was also silent on these opening remarks, which were rhetorically significant as they contained the speech’s emotional ballast or *pathos*. Nonetheless, before we decided that the *Congressional Record* account should be given precedence, we found problems that would probably have gone unnoticed were it not for the CollateX outputs.

In the weeks preceding Parnell’s speech to Congress, the English historian J.A. Froude published a series of articles about Irish history and politics in *The North American Review*. Parnell quoted an extract from one of these articles in which Froude described the land system as the worst of the “fatal gifts” England had bestowed upon Ireland. In the *Congressional Record* the first sentence of the quotation was recorded as, “But – of all the *feudal* gifts which we bestowed upon our unhappy possession was the English system of owning lands”. By contrast, *The Washington Post* reported it as, “But, of all the fatal gifts which we bestowed upon our unhappy possession was the English system of owning land”. *The Irish World* likewise printed the term “fatal gifts”. Leaving aside the differences in punctuation and the plural versus singular of land versus lands, the reports disagree on whether Parnell said “fatal gifts” or “feudal gifts”.

The inconsistency is not significant in its own right, as the disagreement most likely comes from a congressional stenographer mishearing the speaker or from a slip of the tongue on Parnell’s part. It is certainly possible that Parnell misspoke and *The Washington Post* and *The Irish World* corrected his mistake while the *Record* did not. Yet, given the prestige and authority of the *Congressional Record* as the record of the United States Congress, such mistakes have a legacy in the garbled transmission of Parnell’s speech. For example, a volume published in approximately 1904 entitled *Irish Literature* printed an edited version of the *Congressional Record*’s account of Parnell’s speech and the volume’s editor, the writer and politician Justin McCarthy, failed to correct the error. Additionally, in a 1986 debate regarding the recently signed Anglo-Irish Agreement and the responsibility of the United States for helping to secure peace in Northern Ireland, Senator D.P. Moynihan commended Parnell’s speech to the Senate, leading to its republication in the record of that body, replete with the misquotation of Froude.

*The Irish World*’s report deserves its own discussion, as it bucked the trend by demonstrating significant editorial intervention in its reporting of the address to the House. Unlike the myriad other variations in the dataset, the divergences in the *World*’s report display signs that they were ideologically motivated. *The Irish World*, edited by Patrick Ford, was widely read on both sides of the Atlantic, circulating over 60,000 copies in the United States with a further 20,000 copies circulating in Ireland and Great Britain [Dungan 2014]. As none of the major Irish newspapers printed a full report, Ford’s newspaper was the context in which most Irish readers would have encountered Parnell’s address. Furthermore, during the 1888-89 special commission into Parnellism and crime, one of Parnell’s counsels, H.H. Asquith, read *The Irish World*’s report into the record as an authoritative account of Parnell’s speech (*Special Commission Act, 1888*). Its role in the afterlife and reformulation of the speech grants it a status that outweighs its unreliability as evidence for what Parnell actually said before the House.

*The Irish World* dotted sub-headings, such as “REPLACE THE ARTIFICIAL BY THE NATURAL” and “CONDEMNED BY ENGLISH AUTHORITY”, intended to guide the reader through the speech. The paper also capitalised and italicised certain words and phrases to enhance their effect on the reader. However, the most significant variations in *The Irish World*’s account were the notable silences it contained in relation to the other two accounts. In the passage that the *World* labelled “THE OVERPOPULATION TALE”, large portions dealing with emigration and overcrowding were expurgated:

parnell_source_00382 <i>Congressional Record</i>	parnell_source_00386 <i>The Washington Post</i>	parnell_source_00633 <i>The Irish World and American Industrial Liberator</i>
Now, we have been told by the landlord party	Now, we have been told by the landlord party	Now, we have been told by the landlord party
	,	,
as their defense of this system	as their defense of this system	as their defense of this system
	,	,
that the true cause of Irish poverty and discontent is the crowded state of that country	that the true cause of Irish poverty and discontent is the crowded state of that country	that the true cause of Irish poverty and discontent is the crowded state of that country
,	,	
and the only remedy emigration;		
and I admit to the fullest extent that there are portions of Ireland which are too crowded. The barren	and I admit to the fullest extent that there are portions of Ireland which are too crowded. The barren	
hills	lands	
of the west of Ireland, whither the people were driven from the fertile lands after the famine, are too crowded	of the west of Ireland, whither the people were driven from the fertile lands after the famine, are too crowded	
;	,	,
but the fertile portions of Ireland maintain scarcely any population at all, and remain as vast hunting-grounds for the pleasure of the landlord class. Before	but the fertile portions of Ireland maintain scarcely any population at all, and remain as vast hunting-grounds for the pleasure of the landlord class. Before	but the fertile portions of Ireland maintain scarcely any population at all, and remain as vast hunting-grounds for the pleasure of the landlord class. Before

Table 4. Collatex output for Parnell’s speech to the House of Representatives, January 1880

Two sentences further on we find another major omission. The *Congressional Record* and *The Washington Post* reported in terms closely resembling each other that Parnell declared:

Let the next emigration be from the West to the East, instead of from the East to the West – from the hills of Connemara back to the fertile lands of Meath. When the resources of my country have been fully taken advantage of and developed, when the agricultural prosperity of Ireland has been secured, then if we have any surplus population we shall cheerfully give it to this great country. Then our emigrants will go willingly and as free men – not shoved out by a forced emigration, a disgrace to the Government whence they could come and to humanity in general. [Applause.] Then our emigrants would come to you as come the Germans, with money in their pockets, and education to enable them to obtain a good start in this great and free country, with sufficient means to enable them to push out to your western lands, instead of remaining about the eastern cities, doomed to hard manual labor, and many of them falling a prey to the worst evils of modern city civilization.

49

*The Irish World*, however, skipped this passage altogether. The image of Irish immigrants to the United States that Parnell painted was not a flattering one. Considering the nature of *The Irish World's* predominantly urban, immigrant Irish-American readership, it is possible that the *World* did not wish to show Parnell as having said words to this effect.

50

The last significant omission in the *World's* report was to do with Parnell's use of land reform in Prussia as a model for reform in Ireland. In an open letter to Parnell, *The Irish World* urged him not to make comparisons with European land systems but, as Paul Bew phrased it, Parnell "pointedly" flouted this recommendation [Bew 1979]. In the open letter, the *World* declared that people who use examples from the European continent to argue for changes to land tenure in Ireland were "half-way men". Parnell's use of the Prussian example placed him out of ideological alignment with the *World*. In return, the *World* excised most of his discussion of land reform in Prussia.

51

## 6.3 Tools

The SBERT Sentence Transformers and sklearn Libraruiues were easy to use and both in establishing a general level of similarity between sources, which proved crucial to the clustering process, and in elucidating general patterns in the reporting of Parnell's speeches. Experimenting with the distance threshold to find an appropriate level of similarity was an essential part of this process.

52

The Collatex software proved very effective at collating texts which had a sufficient level of general similarity. With texts with low levels of similarity the software collated on false positives generated by commonly used words or punctuation. For a corpus containing texts with very variable levels of similarity, a preprocessing stage to establish which groups of texts are suitable for collation becomes essential. As pointed out by the developers of the library, Collatex (as with all collation software) sometimes struggles with transposed text. This is a known problem which the developers are working on. The outputs produced by the software can be slightly confusing to read, especially when run across multiple texts, as described above in Section 5.3.

53

## 7 Conclusions

The implementation of automated collation on our corpus brought with it a number of challenges. Whilst we had some success in building a workflow for collation, this necessarily involved a sub-process of finding sources that were similar enough for the automated tools to pick out similarities and differences in a coherent manner. When speech reports, or parts of reports, were too dissimilar, the collation outputs devolved into picking out false positives, such as isolated instances of punctuation. The SBERT library was extremely useful in enabling us to find clusters of reports around speeches that were similar enough to be collated. This process also highlighted significant variations in journalistic practice, such as the fact that Ulster Unionist newspapers tended only to report what Parnell said in summary.

54

As a research tool, our collations enabled us to highlight discrepancies between sources more easily, drawing attention to additions and omissions as they occurred. This allowed for notable insights, making us aware of errors in respected sources such as the *Congressional Record* that have gone on to be quoted in other contexts. With the exception of *The Irish World*, we were unable to see the influence of ideological bias in the ways different sources reported the speeches, but this is an insight in itself, perhaps revealing a general editorial trend towards objectivity in relation to speech reports or the reliance of many publications on a single eyewitness account. Ideological bias seemed to be mainly expressed through other journalistic practices, such as only providing short summaries of speeches, or omitting to report on them altogether.

55

In terms of new information, the project told us less about Parnell and his speeches, and more about the practices of the publications, editors and journalists who reported (or failed to report) on them. This concentration on the factors which went into the creation of reports of the speeches will help us to see Parnell and the reception of his ideas in a broader and richer context. As regards the methods themselves, the project highlighted the potential of digital approaches to raise new questions and therefore generate new research pathways. It was fascinating and very fruitful to see the interplay of digital method and scholarly expertise in practice, as the results of the methods generated new insights for the project team, which then fed back into the methodology in a highly productive iterative process.

56

Building on the work described in this article, project members have submitted a funding proposal to construct a "transparent edition" of Parnell's speeches, including a full and open publication of the TEI dataset which forms the basis of the edition (both as data and through a simple web interface), open and documented code for all of the tools used or developed in the course of creating the edition, and the critical edition itself. By doing this we hope to make explicit the iterative and exploratory relationship between digital methods and editorial work which has proved so fruitful in understanding Parnell's speeches and their context.

57

## Notes

[1] The Boston Irish-American newspaper *The Pilot* printed a full report that was credited to the *Congressional Record* and which was almost identical to the *Record's* account; see *The Pilot*, 14 Feb. 1880. After a delay of two weeks, *The Freeman's Journal* printed a report that was almost identical to the *Post's*; see *Freeman's Journal*, 16 Feb. 1880.

## Works Cited

- Bew 1979** Bew, P. (1979) *Land and the National Question, 1858-82*. Atlantic Highlands, NJ: Gill and MacMillan.
- Bew 2012** Bew, P. (2012) *Enigma: A New Life of Charles Stewart Parnell*. Dublin: Gill Books.
- Birnbaum and Spadini 2020** Birnbaum, D. and Spadini, E. (2020) "Reassessing the locus of normalization in machine-assisted collation", *Digital Humanities Quarterly*, 14(3).
- Boyce and O'Day 1991** Boyce, D. and O'Day, A. (eds) (1991) *Parnell in Perspective*. London: Routledge.
- Cummings 2019a** Cummings, J. (2019a) "Opening the book: Data models and distractions in digital scholarly editing", *International Journal of Digital Humanities*, 1, pp. 179–93, <https://doi.org/10.1007/s42803-019-00016-6>.
- Cummings 2019b** Cummings, J. (2019b) *A world of difference: Myths and misconceptions about the TEI*, *Digital Scholarship in the Humanities*, 34 (Supplement 1), pp. 58–79, <https://doi.org/10.1093/lc/fqy071>.
- Dekker et al. 2014** Dekker, R., Van Hulle, D., Middell, G., Neyt, V., Van Zundert, J. (2014) "Computer-supported collation of modern manuscripts: CollateX and the Beckett Digital Manuscript Project", *Digital Scholarship in the Humanities*, 30(3), pp. 452–70, <https://doi.org/10.1093/lc/fqu007>.
- Dungan 2014** Dungan, M. (2014) *Mr. Parnell's Rottweiler: Censorship and the United Ireland Newspaper, 1881-1891*. Dublin: Irish Academic Press.
- Houghton et al. 2020** Houghton, H., Parker, D., Robinson, P., Wachtel, K. (2020) "The *Editio Critica Maior* of the Greek New Testament: Twenty years of digital collaboration", *Early Christianity*, 11(1), pp. 97–117, <https://doi.org/10.1628/ec-2020-0009>.
- Jackson 2003** Jackson, A. (2003) *Home Rule: An Irish History, 1800-2000*. London: Weidenfeld & Nicolson.
- Jackson 2012** Jackson, A. (2012) *The Two Unions: Ireland, Scotland, and the Survival of the United Kingdom, 1707-2007*. Oxford: Oxford University Press.
- Low 1991** Low, D.A. (1991) *Eclipse of Empire*. Cambridge: Cambridge University Press.

- Matthew 1987** Matthew, H.C.G. (1987) "Rhetoric and politics in Britain, 1860–1950", in Waller, P.J. (ed.) *Politics and Social Change in Modern Britain*. Brighton: Harvester, pp. 34–58.
- McCarty 2013** McCarty, W. (2013) "Knowing: Modelling in literary studies", in Siemens, R. and Schriebman, S. (eds.) *A Companion to Digital Literary Studies*. London: Blackwell, pp. 391-401.
- McGann 2022** McGann, J. (2022) "Editing and curating online: Beginning again", *Textual Cultures*, 15(1), 53-62.
- Nabugodi and Ohge 2022** Nabugodi, M. and Ohge, C. (2022) "Provocations towards creative critical editing", *Textual Cultures*, 15(1), pp.1-10.
- Travers 2000/2001** Travers, P. (2000/2001) "Reading between the lines: The political speeches of Charles Stewart Parnell", *Studia Hibernica*, 2000/2001, 31, pp. 243-256.
- Travers and McCartney 2013** Travers, P. and McCartney, D. (eds.) (2013) *Parnell Reconsidered*. Dublin: University College Dublin Press.
- Van Hulle 2019** Van Hulle, D. (2019) "Artificial imagination, imagine: New developments in digital scholarly editing", *International Journal of Digital Humanities*, 1, pp. 137–40, <https://doi.org/10.1007/s42803-019-00020-w>.



This work is licensed under a Creative Commons Attribution-NoDerivatives 4.0 International License.