WIENER digitale REVUE

Zeitschrift für Germanistik und Gegenwart

Vincent Neyt und Dirk Van Hulle

Developing the *Beckett Digital Manuscript Project*

DOI: 10.25365/wdr-01-03-01

Lizenz:
For this publication, a Creative Commons Attribution 4.0 International license has been granted by the author(s), who retain full copyright.

Vincent Neyt and Dirk Van Hulle

Developing the *Beckett Digital Manuscript Project*

Introduction

1 Samuel Beckett's manuscripts are scattered over more than a dozen archives on both sides of the Atlantic. To study them, scholars needed to travel to several places, because the manuscripts of, say, *Krapp's Last Tape* are dispersed over various holding libraries. That is why we try to digitally reunite all of Beckett's manuscripts. The *Beckett Digital Manuscript Project* (BDMP) offers genetic digital editions of Beckett's works, along with a reconstruction of his personal library. Because of copyright restrictions, these resources require a subscription fee on an individual or institutional basis. There is also a "free zone", offering a demo of one of the genetic editions and of the *Beckett Digital Library* (BDL), extensive documentation and several "sneak peeks" into the material through notebook thumbnails, statistics on the number of additions and deletions within all genetic documents stored in the editions, and most recently a first installment of Breon Mitchell's *Samuel Beckett: A Bibliography: Part I: The Early Years: 1929–1950*.

Genetic editions

2 At the moment of writing, there are eight genetic editions in the BDMP: *Molloy*, *Malone Meurt / Malone Dies*, *L'Innommable / The Unnamable*, *En Attendant Godot / Waiting for Godot*, *Fin de partie / Endgame*, *Krapp's Last Tape / La dernière bande*, *Stirrings Still / Soubresauts and Comment dire / what is the word.*

3 The point of entry for each genetic edition is a catalogue of all documents that are part of the work's genesis, in chronological order. The documents come from archives all over the world, such as the University of Reading, the Harry Ransom Center in Austin, Texas, Trinity College, Dublin and the Bibliothèque nationale de France. The *Fin de partie / Endgame* module, for instance, brings together documents from five different partner institutions: the University of Reading, Trinity College Dublin, The Harry Ransom Center, Ohio State University and Harvard University.

4 For each of the documents, its name, the description of the archive's catalogue and a thumbnail of the cover or first page is given. Even without a subscription, the catalogue of the genetic editions can be browsed. The Manuscript Chronology puts all of the documents on a timeline and indicates the relations between versions.

Browsing

5 Every document has been scanned at high resolution and transcribed in XML, following the P5 guidelines of the Text Encoding Initiative (TEI). A document can be browsed in several ways. An "About" page offers more metadata, a thumbnail of every page, and links to the different "views" the BDMP offers: the text view, the image view, the image/text view, and the TEI/XML view.

6 The text view presents the transcription as a running text. In the case of notebooks, the verso and recto pages are put side by side. This view caters for users who want to read the text of a document. It is no longer in Beckett's sometimes near-illegible handwriting, and has been linearized: additions in the margins and on the facing leaf have been inserted at the appropriate point in the text.

7 The images can be viewed in a number of ways: one image for each page, with or without a zoom lens; as two facing pages in the case of notebooks (the "double page view"); and lastly the image/text view, where the image is divided into zones and the transcription can be called up at any time by clicking on a zone.

Compare sentences / CollateX

8 In the transcriptions, all the sentences have been tagged and given a number, based on a numbering of the sentences in the first edition of the work. When the user is in the image/text view or the text view, the option "Compare sentences" appears in the navigation bar. Toggling this option makes the sentence numbers appear at the beginning of each sentence in the transcription. The links lead to the "Synoptic Sentence View": all the versions of that sentence are listed chronologically. Loose jottings ("paralipomena") that relate to the sentence but are not versions in the strictest sense, are also included.

9 The BDMP has not encoded a critical apparatus of the variation between the sentence versions. Instead, it has incorporated the automatic collation software CollateX[1] into its architecture. The sentences are sent to our CollateX instance via its REST service, and the output from the program is visualised as an alignment table. Text that remains unchanged through the versions is placed in columns with a green border, and differences are highlighted in columns with a black border and gray background color.

10 CollateX allows additional information, such as the indication whether something is a "deletion" or an "addition", to be passed along the collation process, allowing us to visualise deletions and additions in the alignment table in the same way as in the regular transcriptions: deleted text struck through and added text in superscript.

Doodles

11 Beckett typically left the verso pages in his notebooks blank and wrote only on the recto side. The versos were meant to house additions and notes. Many of these versos contain doodles. In the XML, doodles are tagged with a <figure> element. At the time of writing, there are 1526 of these doodles in our corpus. As a way of describing them, the editorial board of the BDMP put together a "doodle categorization". The four main categories are *object*, *organism*, *shape* and *symbol*. All four contain several subcategories. These categories can be called up by clicking on the doodle in the text and image/text views, and can be searched and sorted through the search engine (see below: Suggested searches). If there is a clear relation between a doodle and the text on the (facing) page, the editor has added a description explaining the link.

The Search Function and Suggested Searches

12 A search box is permanently visible in the navigation bar. Users can perform full-text searches on all documents in all genetic editions and in the *Beckett Digital Library* (BDL). Both the transcriptions and the editorial annotations are searchable. Searches can also be limited to a specific work, to transcriptions or doodle categories only; they can also be filtered by work and by document.

13 To draw attention to certain interesting phenomena in the texts, we have put together a list of suggested searches. Users can retrieve all *doodles* and *diagrams*, and refine the results by selecting individual doodle categories. As the number of doodles keeps growing, this feature provides a much-needed way of finding a doodle based on what it depicts, or its shape, or a combination of shapes. Other suggested searches focus on *calculations*, *addresses* and *phone numbers*, *stage drawings* and *musical scores*. We have tagged *intertextual references* in the transcriptions, and all of these can also be called up together through the suggested search menu. As Beckett often marked the date at the start of a writing session, a search for all of these dates in all of the documents can also be called up from this menu. A last option in the list is gaps: places where Beckett left a blank space to be filled in later.

Beckett Digital Library

14 In 2016, we incorporated the *Beckett Digital Library* in the BDMP: a digital reconstruction of Samuel Beckett's personal library, based on the volumes preserved at his apartment in Paris, in archives (such as the *Special Collections* of the University of Reading) and private collections (James and Elizabeth Knowlson Collection, Anne Atik, Noga Arikha, Terence Killeen, ...). It currently houses 762 extant volumes, as well as 247 virtual entries for which no physical copy has been preserved. We have transcribed all reading traces, marked words and marginalia. Beckett often used the books that he read in his own writing. Whenever a passage is alluded to in the drafts of his works, this is highlighted by a "Manuscript Link" in the bibliographic description of the volume in question. A link is also present in the transcription of the fragment in the draft. A clickable "L" icon provides some information and a link to the BDL.

15 After this general introduction to the BDMP, we will address some more advanced features.

TEI Encoding: Text vs Document

16 When the Text Encoding Initiative released its 2.0.0 version of the P5 encoding guidelines in 2011, it introduced a major addition. Historically, the TEI has always been "text-oriented", supplying elements to encode textual structures and features. Because the internet has proven to be so well-suited to publishing images of individual manuscript pages, the need arose for "document-oriented" text encoding. The 2.0.0 version of P5 introduced new elements to encode a document as a series of <surface> elements, each <surface> element containing <zone> elements.

17 The BDMP, started in 2008, was "text-oriented" from the start. Nonetheless, we did encode a separate (division) for each page, which also caters for a "document-oriented" approach, that can be visualized in the form of a page-per-page text view and a zone-per-zone visualization in the image/text view. The advantage is that it looks like documentary approach, but is also text-oriented. The rationale behind this

approach is that the integrity of the sentences is crucial. We also need the "text-orientation" to see to it that sentences that are broken off (e.g., at the end of a page) are put back together again to provide options such as "Compare sentences", "CollateX" and the "Search Engine".

Distant reading

18 The BDMP offers two features that would classify as "distant reading". First, under "Free Features", users can access a page called "Writing Sequence of *L'Innommable*". It visualises a hypothesis on how Beckett filled the pages of the two notebooks he used for his first draft of the novel *L'Innommable*, sentence by sentence. Using the BDMP's sentence numbers for the work as a starting point, it (1) shows where each sentence is written (i.e., on which page in which notebook); (2) indicates where each sentence eventually ended up (in the published text); and (3) visualizes an editorial hypothesis on the sequence in which these sentences were written. These goals are achieved by a sequence of five different visualizations, the last of which ("step 5") offers an animated interpretation of the writing sequence.

19 Second, we offer statistics for each of the genetic editions. For all documents in all editions, pie charts show the number of added, deleted and modified words. The first draft of *En Attendant Godot*, for instance, counts 20972 words that remained unmodified during the writing of that draft; 1719 words were struck out; 565 words were added and 19 were modified. In addition to the pie charts for each genetic edition, we also provide two comparisons: (1) a comparison of all first draft versions of all works currently included in the BDMP side by side; (2) a comparison of the total words in the full genetic dossier of all works currently in the BDMP.

20 By making the TEI/XML source file available for each transcribed document, we hope to encourage other researchers to perform distant reading techniques on them. The BDMP materials are relatively modest in size. Most distant reading techniques are applied to corpora of hundreds, if not thousands of texts. However, the texts used are usually not marked up (plain text) and based on uncorrected OCR from mass-digitisation projects. The BDMP transcriptions could provide an interesting challenge to the field, as they are very precise and enriched with additional information: added and deleted words, speakers and speeches (in the case of performance texts), sentence demarcations, and so on.

ModNets

21 Modernist Networks ("ModNets") is a federation of digital projects in the field of modernist literary and cultural studies.[2] It aims to bring together "digital objects" from these projects and currently houses 84,901 peer-reviewed digital objects from 69 federated sites. A "digital object" can be an image, a transcription, metadata, or a combination of these elements. Copyright restrictions prevent us from submitting images and transcriptions, but we do provide metadata for all documents in the genetic editions.

Handwritten Text Recognition

22 An exciting development in the field of Digital Humanities of the last five years is "Handwritten Text Recognition" (HTR), a kind of OCR for handwritten materials as it were. At this moment, the leading project in this area is *Transkribus*.[3] The software allows users to upload images and run a tool to decipher the text of a particular author. For the output of that tool to be useable, the program first needs to learn how to

"read" this author's handwriting. An author algorithm must be trained by supplying test material: images and their corresponding transcripts. In collaboration with the *Transkribus* team, a Beckett algorithm was trained on 400 transcribed pages. In the case of a difficult handwriting such as Beckett's, a high number of test pages is required to achieve a "character error rate" that is low enough for the result to be useable. We also discovered that the algorithm performed better if a separate algorithm was trained per language: there is now a French Beckett algorithm and an English one.

23 *Transkribus* offers a number of export formats, TEI/XML being one of them. We can envisage *Transkribus* becoming part of the workflow for upcoming BDMP genetic editions. An editor uploads the images of a document, uses the HTR tool to produce a first version of the transcription, makes the necessary corrections and adds inline tags such as <add> and <del>, and exports as TEI/XML. Currently the BDMP team is working on a research project called "CATCH 2020: Computer Assisted Transcription of Complex Handwriting" (University of Antwerp), which aims to improve the way in which *Transkribus* handles supralinear and marginal additions and performance texts. The deciphering could also be enhanced by supplying stylometric information about the author and other versions of the same text fragment.

Future Plans

24 In future, the BDMP will grow in two directions: (1) horizontally, as it will incorporate more and more modules, and (2) vertically, as it will develop more tools for users to be able to explore it more profoundly across modules. That said, it is crucial that the technical sophistication does not eclipse the BDMP's large hermeneutic potential for literary scholars. This danger is at least partially remedied by the accompanying BDMP volumes, the printed books that provide what Hans Zeller called the 'Deutung' (interpretation) to the 'Befund' (the record of texts) visualized in the digital format. The complementary nature of the BDMP project (digital modules plus printed volumes) may hopefully contribute to the much-needed synergy between the 'old' and 'new' technologies of scholarly editing.

Notes

1 See https://collatex.net. The BDMP uses version 1.7.1.
2 http://www.modnets.org/about/what-is-modnets/
3 https://transkribus.eu/Transkribus/

Abstract

Dirk Van Hulle and Vincent Neyt report on the Beckett Digital Manuscript Project. Its purpose is to reunite the manuscripts of Samuel Beckett's works in a digital way, and to facilitate genetic research: the project brings together digital facsimiles of documents that are now preserved in different holding libraries, and adds transcriptions of Beckett's manuscripts, tools for bilingual and genetic version comparison, a search engine, and an analysis of the textual genesis of his works.

Zusammenfassung

Der Werkstattbericht von Vincent Neyt und Dirk van Hulle liefert Einblicke in das Beckett Digital Manuscript Project (BDMP), das Samuel Becketts weltweit verstreute Manuskripte im Hypertext vereinen wird. Das Projekt führt auf einzigartigem Weg digitale Faksimiles, Transkriptionen und textgenetische Auszeichnungen zusammen: Nicht umsonst gewann es 2018 den *Modern Language Association Prize for a Bibliography*.

Keywords: Samuel Beckett, Digital Library, Digital Manuscript Project, TEI, Transkribus

Schlagwörter: Samuel Beckett, Digitale Bibliothek, Digital Manuscript Project, TEI, Transkribus

Authors

Vincent Neyt
University of Antwerp

Dirk Van Hulle
University of Antwerp