The Document as Event

Assessing the Value of Digital Collections for Theatrical Heritage

Thomas Crombez

When the performance is over, nothing remains but the memories of the participants and witnesses. At least, that is the well-established cliché in the performing arts, and in theater history as well. To investigate a theatrical event is, by definition, to investigate a historical event. It implies for that matter the study of the event's traces: the material record of costumes, props and set designs; the written scenario or script (when available); and the eyewitness accounts of performers, technicians and spectators.

For theater historians, there is an important extra source: the critical record of reviews and essays. Historical documents of theater criticism can bring a former performance context and its related practices (acting styles, physical communication, audience reactions) back to life. Only recently have such documents become accessible through large-scale digital collections. The following research questions will be dealt with in this contribution. I will examine how reviews can be considered part of theatrical heritage, and how they can be valorized through digital means. What are the future possibilities for digitally expanded reviews? Ideally, what would we want digital collections of reviews to look like? To answer these questions, I will discuss new approaches to theater discourse using tools from the nascent field of digital humanities.

Theatrical Documents as Theatrical Heritage

Cultural heritage is traditionally defined as being composed of both material and intangible heritage. In the case of theater, the 'heritage object' is extremely diverse, as every example of material heritage, such as a theater building or a costume, is intertwined with cultural practices, such as a particular style of acting or audience response. However, even theater scholars tend to prioritize material traces of theatrical architecture or dramatic practice over intangible ones, as the former are more readily available and accessible than subjective accounts of acting styles. Reconstruction projects, such as the neo-Elizabethan Globe in London, have a privileged claim to authenticity on the mere basis of their very tangibility.[1] As Timothy De Paepe pointed out, "a real building, or even a physical scale model can give a false sense of comfort, of certainty."[2]

Taken in itself, the theatrical review seems a relatively low-value form of material heritage. It is regarded as a merely verbal and subjective account of a theatrical event. Yet, when considered in larger numbers, the review may constitute a crucial source of intangible heritage. Newspapers and specialized press provide an enormous amount of evidence to modern theater history. The sheer abundance of theater reviews from the eighteenth century onward ensures that theatrical reviews provide invaluable sources for imagining and examining theatrical practice. This requires, however, that the wealth of information they hold about plays, playwrights, actors and companies is adequately and efficiently retrieved—or, as contemporary parlance would have it, 'mined'—from the texts.

Digital Collections of Documents

Notwithstanding their profusion, theatrical documents often prove difficult to access. The contemporary researcher or enthusiast who wants to consult theatrical reviews must generally do so by browsing through piles of bound volumes of periodicals in a library. This time-consuming approach is feasible for individual case

[1] Conkie 2006.

[2] De Paepe 2010, 32.

studies, such as the examination of an individual dramatist or playhouse within a specific timeframe, but for any broader kind of inquiry the method quickly reaches its limits.

Digital collections provide an excellent solution to the proverbial 'needle in a haystack' problem. A properly indexed digital collection makes documents not merely browsable, but fully searchable. Over the last few years, a number of noteworthy large-scale digitization projects of periodicals have been launched. One of the showcase projects of *Google Books*, the mass digitization project started in 2004, is the fully browsable archive of back issues of *LIFE Magazine* (1936-1972).[3] A second example is the *Historische Kranten* ('Historical Newspapers') project, which was started by the Nederlandse Koninklijke Bibliotheek (Dutch Royal Library) in 2006 and has made available more than eight million pages from Dutch-language newspapers from the seventeenth century to 1995.[4] Less recommendable is the newspaper digitization initiative of the Belgian Koninklijke Bibliotheek 'Albert I' (Royal Library). Although the institution digitized c. 3.2 million pages from seventy periodicals of the nineteenth and twentieth century, it is only possible to consult this invaluable resource through "five special PCs in the reading room."[5]

Such new resources obviously have a huge impact on ongoing and future historical research. Cultural historians will in particular benefit from the new availability of sources. Still, most online newspaper databases are conceived for a broad audience, and hence feature a general-purpose interface for browsing and searching. For instance, a search query in *Historische Kranten* for articles from the interwar period featuring "Amsterdam Stadsschouwburg" in their text returns more than 31,000 results, each of which can be viewed within the context of the original newspaper page.[6] Unfortunately, the theater historian cannot expect any extra help from the website in filtering the articles according to the theatrical events they announce, review or advertize. Similarly, *Google* uses basic tools from computational linguistics to determine which words and expressions are characteristic for the book or magazine issue that the user is currently browsing. Take, for instance, the February 1937 issue of *LIFE Magazine*: the *Google* interface is happy to inform me that "Leon Trotsky," "Reichstag," "Studebaker" and "Kleenex" are among its most common terms and phrases, but it is obviously blind to the question whether these names belong

[3] The digital archive of *LIFE Magazine* is available at: <http://books.google.be/books/about/LIFE.html?id=NoEEAAAAMBAJ> (last accessed 11 April 2014).

[4] See <http://kranten.delpher.nl> (last accessed 11 November 2014).

[5] See <http://www.kbr.be/collections/journaux/journaux_nl.html> (last accessed 11 April 2014).

[6] Search query executed 16 June 2013 on <http://kranten.kb.nl/> using 'Advanced Search' (period: 1919 to 1939; search term: "Amsterdam Stadsschouwburg").

to people, organizations, places or brands.[7] Neither does the search engine provide an orderly and meaningful categorization of these expressions.

Digital archives for that matter create at least as many problems as they solve. Language technology has become so pervasive in everyday life that users have come to foster very specific, but also quite restricted, expectations concerning searching and browsing. They prefer to direct their questions to a simple search box, such as the *Google* search interface. The limitations of this model are evident in the recent efforts that *Google* itself has made to enrich search results with semantic metadata. On the current search results page, a number of options in the top or sidebar of the screen allow users to narrow down the results according to a certain domain or type of results (*e.g.*, news, images or books). Videos, photos and maps (if relevant) identify the search terms in a visual way. In 2012, the internet company introduced the *Knowledge Graph*, which combines a wide-ranging database with thorough analysis of search queries in order to provide more 'semantic,' *i.e.* meaningful, information to users' queries. The system, for instance, now recognizes when a user is looking for a well-known name, such as "Leonardo da Vinci." It will try to provide not merely web pages in which that name is mentioned (*unstructured* text), but it will also give as much *structured* information as available. In the case of "Leonardo da Vinci," the *Google* interface will present information on the artist's biography (date of birth, place of birth, occupation, place of death), family and works.[8]

Semantic enrichment of data is also the approach commonly adopted by scholarly digitization projects. In contrast to mass digitization, research projects such as the Rosetti Archive aim to provide more metadata, but they are necessarily much more limited in scope.[9] Such projects do not rely on automatic enrichment but depend on editors who have *manually* marked up the digitized texts, adding semantic labels to certain documents, passages or expressions. For example, the *Corpus Toneelkritiek Interbellum* ('Corpus of Theater Critiques from the Interwar Period'), which I developed and launched in 2008, contains seven hundred Flemish theater reviews from the period 1919-1939.[10] For each document, the interface includes full performance data of the production under review, detailing the title of the work, dramatist, director and company.

[7] See the homepage of this issue of *LIFE Magazine*: <http://books.google.be/books?id=VFEEAAAAMBAJ> (last accessed 11 April 2014).

[8] Singhal 2012.

[9] See <http://www.rossettiarchive.org/> (last accessed 11 April 2014).

[10] See Crombez 2008.

A digital corpus with rich metadata may be constructed manually when dealing with a limited set of documents. For large-scale projects, by contrast, this method would be unfeasible—it is precisely these laborious efforts of the theater researcher that stand in need of automation. Given the emergence of digital corpora of historical periodicals, could periodicals be mined for theater-related information? Can we imagine and develop software that *detects theatrical events* in vast collections of text? In order to answer these ambitious questions, I would first like to take a closer look at the concepts of *event* and *document*.

From the Theatrical Event to the Document as Event

Before a theater review becomes the document of a theatrical practice, it is first a document about a theatrical *event*. Here I will use the concept of the event as an opportunity to re-think the nature of the theatrical review in its capacity as a document. In other words, I would like to conceptualize *the document as event*, in order to transfer something of the dynamics of the event onto the document, which is commonly conceived of in a rather static fashion.

A review is occasioned by a theatrical event. However, *as a document, it can also be considered to be an 'event' in itself*. The text functions as a linguistic meeting space for a wide diversity of named entities. These include the names of dramatists, performers and directors, of theaters, companies and schools, and the titles of productions and works. In this sense, the totality of all theater reviews ever published (dispersed over countless periodicals, many of which are probably already lost) could be said to hold a virtual record of modern theater history. Even if only a small fraction of this textual treasure could be mined, it might produce a gargantuan database of past theatrical events.

Such an ambitious text-mining project is obviously difficult to realize using current scholarly resources. The chief difficulty lies in having a sufficiently large set of digitized historical reviews at one's disposal. Such collections, as I have shown above, are currently uncommon and, if available, not always very accessible. Still, given a large-scale digital corpus, it is possible and feasible to use off-the-shelf technology from the field of natural language processing (NLP), and more particularly *information extraction*, in order to 1) extract named entities from the digital text, and 2) detect meaningful relationships between those entities.[11] These relationships will often be specified in the document text, but they may also be implicit and presupposed. For instance, internet companies such as *Google* have developed data sets of relations about public figures based on pages from the English-language *Wikipedia*. Recently, *Google* released a data set of 50,000 such relationships, each of which had been verified by five human assessors. For example, 10,000 of these instances detail a relationship of 'place of birth' between the name of a person and the name of a place.[12]

Text-mining could thus be used as a tool for building a detailed database of theatrical events, provided that a digital corpus of historical theatrical documents be made available beforehand. It is my conviction that this is not merely a chimerical vision. Still, leaving aside the conditional tense that has so far dominated this essay, what can be done right now? In the following section of this contribution, I will show how an existing digital collection of late twentieth-century theatrical documents can be semantically enriched using a technique called *supervised machine learning*.

A Case Study of Semi-Automatic Semantic Enrichment

Which techniques can be used for automatically enriching the contents of a digitized collection of theatrical documents with extensive metadata? That such metadata may be of crucial importance to theater scholars has become evident from the above. Semantically enriched data would allow searches for certain terms not merely in a verbal or linguistic sense, but additionally *according to the role* they fulfill in the text. For example, it might become possible to query the Dutch newspaper collection *Historische Kranten* for a certain theater, say, the Amsterdam Stadsschouwburg, and then ask the system to select all names from those articles that occur in the role of actor/actress. The result would be, approximately, the full group of performers that have played at the Stadsschouwburg in the period under review.

Two techniques from natural language processing and machine learning are crucial in the process of transforming plain-text documents into semantically

[11] These two tasks are commonly designated as *named entity extraction* and *relation extraction*, see Jurafsky and Martin 2009, chapter 22 ('Information Extraction'); Hobbs and Riloff 2010.

[12] Orr 2013.

marked-up data. First, all named entities should be detected and extracted from the text, preferably according to their type (personal name, organization, place name). Second, the extracted names have to be classified according to their specific role. In the example below, I will focus on detecting *artistic roles of personal names* in the context of the performing arts (*i.e.*, roles such as director, performer, or critic). It is evident that the same methodology could also be used for distinguishing other names, such as organizational names (with roles such as: 'theater companies,' 'playhouses' and 'funding bodies').

The digital corpus in question concerns the Flemish magazine *Etcetera*, which was started in 1983 in order to track innovative tendencies in the Dutch and Flemish performing arts—the periodical has been instrumental in launching the international careers of artists such as Jan Fabre, Ivo van Hove, Luk Perceval or Jan Lauwers. In 2011, the back issues of *Etcetera* for the period of 1983 to 2008 were fully digitized and made searchable by the *Platform for Digital Humanities*, a project managed and developed by the author at the University of Antwerp.[13] The digital collection currently comprises 114 issues, containing more than 2,500 articles by 657 different authors. All pages were scanned, converted to text with OCR (Optical Character Recognition) software and manually corrected. In total, the corpus holds more than 5,000,000 words. Named entities (people, places and organizations) were marked up automatically in the texts using Named Entity Recognition software for Dutch that was generously made available by the Language and Translation Technology Team (LT^3) at Ghent University.[14] These additional metadata are presented to the user in the form of a sidebar, which also contains the main metadata (author, date, issue) and a selection of automatically generated keywords.

How does one determine *semantic roles* for the names extracted? In the following paragraphs, I will detail a semi-automatic method of semantic enrichment, the results of which will be evaluated on the basis of how they might enrich the web archive as it is currently online. To conclude, I will make suggestions for improvement of the outcome and mention its potential benefits to researchers.

My approach depends on *supervised machine learning*, which is a subdiscipline of artificial intelligence used in a wide variety of everyday technologies, such as *spam filtering*.[15] Using an existing training set of spam e-mail messages and regular non-spam messages, a spam filter 'learns' which characteristic words or other linguistic patterns distinguish a spam message from a regular message (for instance, a high

[13] See <http://theater.uantwerpen.be/etc> (last accessed 11 April 2014).

[14] Schuurman, Hoste & Monachesi 2010.

[15] Bird, Klein & Loper 2009, 221-260.

incidence of words such as “casino” or “Viagra”). In their most basic form, spam filters and other machine learning applications are built on a statistical technique known as *Bayesian learning*, after Thomas Bayes (1702-1761), the mathematician who first proposed a formula for dealing with conditional probabilities. Given the probability of a prior event, the formula allows one to compute the probability of other events depending on that event. In the case of a spam filter, the prior probabilities concern the chance that a given word will occur (or not) in a spam e-mail. The filter thus decides what probability to assign to the message being either spam or non-spam, depending on the incidences of individual words in spam and non-spam messages.

In the case of identifying artistic roles, the task also entails classification. It would be desirable to classify personal names according to eight types of artistic roles: choreographer, critic, dramatist, photographer, performer (*i.e.*, actor/actress), director, scenographer and theorist. Just as effective spam filtering depends on a large data set of previously labeled spam and non-spam messages, this task also necessitates a database of previously classified names. Therefore, I manually labeled 458 names. All labels are more or less evenly distributed, although the data set is skewed towards the most frequently mentioned role, namely that of director (112 names).

Next comes the more arduous task of *feature selection*. This step determines which linguistic features will be used to classify a given name. Obviously, it is impossible to use the full document (as a spam filter does), since we are not classifying the document as such, but only one of the names occurring in it. An evident technique is to use only ‘neighboring words’: the words that are mentioned closest to the name, for instance in a five-word window before and after the name. It may reasonably be supposed that this window of words will hold more information about the name in question than the text as a whole. An experiment was set up to test if the hypothesis would hold true. I extracted text snippets of five words before and after each occurrence of the names. These windows of ten words (in total) are taken as the ‘features’ to feed to the machine learning algorithm. The presence or absence of certain words in each window will enhance or diminish the probability of the name in the center of the window belonging to one of the eight labels.

All classification tasks require that the data set be divided into a *training set* and a *test set*. From the latter the machine learning algorithm ‘learns’ which words are more likely to occur close to the name of a director (or actor or writer or ...) and which words not. The classifier algorithm is, in other words, first trained on the features and labels of the training set. Then its accuracy is tested on the (unseen) test set. However, even a randomized data set may result in an uneven distribution of labels over both sets. If, for instance, an unusually high ratio of the items in the test set is labeled as choreographer, while the training set holds a correspondingly lower ratio of such items, the accuracy reported on the test set will be unexpectedly low. To counter this, I apply a technique known as *ten-fold cross-validation*. First, the total data set was divided into ten 'folds' of equal size. The software is trained on the features and labels of ninety percent of the data set, and then trained on the remaining (unseen) ten percent. Then, the same procedure is repeated for each fold. Each test run will yield an accuracy score, and the final result is the average of this ten-score set.

Using the Bayesian machine learning module of the *Natural Language Toolkit* (NLTK), the results of the first experiment were positive. The classifier could correctly predict the label of an unseen item in 47.6 percent of the cases. This is significantly higher than the minimal outcome to be expected from a classifier that would simply pick one of the eight labels at random (12.5 percent). Therefore, the preliminary conclusion is justified that a window of ten neighboring words is efficient at predicting the artistic role of a name.

Still, a system that predicts the correct label in less than half of the cases, and hence selects an *inaccurate* label in all other cases, would be unfit for a scholarly resource. In real-world situations, users will expect near-perfect accuracy, or else they will no longer trust the resource and stop using it. The greatest obstacle to real-world application in the setup of the first experiment is the task itself. With an accuracy score of 47.6 percent, the classifier may be said to be successful because the expected score is so much lower. Therefore, if the task is simplified, the accuracy score may climb to a more acceptable level of performance.

The next series of experiments switches from an eight-label classification task to a *binary* classification task. Instead of expecting the classifier to make an accurate selection from a set of eight labels, the task is split up into eight binary subtasks. For each label, a classifier is trained that determines whether or not the words surrounding a name allow them to be assigned that label. In other words, eight successive classifiers determine for each name whether it may relate to a choreographer (or not), a critic (or not) and so on. This modification effectively renders the task much simpler. The expected baseline score against which to judge the classifier's accuracy rises to fifty percent if a label be assigned randomly, compared to just 12.5 percent on the previous task.

The result of the second series of experiments is also positive. Using the same parameters as above (*i.e.*, ten-word snippets), the accuracy scores of the eight distinctive classifiers vary between a minimum of 72.7 percent (for directors) and a maximum of 84.9 percent (for photographers), with an average score of 78.6 percent. In nearly four out of five cases, the system is able to deduce the artistic role of a given name from a ten-word window around that name. All the same, it is not certain that this is the highest achievable accuracy, for what is the optimal size of the text snippets? Perhaps the accuracy increases with longer (or shorter) snippets. And at what point are the neighboring words no longer relevant (on average) for gauging the artistic profession of the person mentioned?

In order to answer these questions, I repeated the experiment with different parameters. First, I reduced the window to three words before and after each name, which led to a slight decrease in accuracy (77.9 percent). Next the text windows were increased to ten words before and after, which caused a small increase in performance (79.3 percent).

To conclude the experiment, I introduced a new feature: *word position.* From previous experiments and closer inspection of the corpus, it became evident that word position might positively influence automatic classification. For instance, photographers' names are regularly mentioned in image captions, immediately preceded by the text ("Foto:"). If word position could also be captured and translated into a feature, the presence of the word "foto" ("photo") immediately preceding the name would probably be a critical parameter for deciding whether a given name was a photographer. The effect proved less noticeable than anticipated: compared to the previous highest accuracy score of 79.3 percent, only a slight increase to 81.1 percent was observed.

Role	**# names in data set**	**# snippets per role**	**6 words**	**10 words**	**20 words**	**20 words (including word position)**
Choreographer	45	4483	0.789	0.809	0.833	**0.844**
Critic	39	4226	0.791	0.776	0.772	**0.797**
Dramatist	51	4524	0.749	0.775	0.763	**0.785**
Photographer	43	782	0.852	0.849	0.863	**0.878**
Performer	66	2979	0.759	0.767	0.786	**0.803**
Director	112	11126	0.726	0.727	0.729	**0.756**
Scenographer	55	412	0.802	0.805	0.802	**0.807**
Theorist	47	2261	0.763	0.782	0.799	**0.815**
Average			0.779	0.786	0.793	0.811

Table 12. Results of text-mining performed in the Etcetera database.

Table 12 summarizes the results described above. The second column details the number of names that were manually selected and labeled for each of the artistic roles under consideration. The third column numbers the snippets retrieved from the whole of the text corpus for each role. All remaining columns give the accuracy scores for the respective experiments, using six-word windows (three before and after), ten-word, twenty-word, and twenty-word including word position.

User Reactions to Automatically Generated Metadata

Is it possible to generalize the results of these experiments and apply the classifier to *all* personal names in the collection of articles? A number of caveats pertain to this conclusion. First, applying all eight classifiers to the whole corpus might result in multiple positive labels. What if two or more labels (for instance, director *and* dramatist) are assigned to one and the same person? The classification might be correct for some figures from theater history—think of Shakespeare, Molière or Brecht. Still, the accuracy of the classifier is only around eighty percent, meaning that it is *incorrect* in the twenty percent of remaining cases. Hence, one of the multiple labels may simply be the result of incorrect tagging. If five occurrences of a given name are labeled as "not a choreographer," and just six as "choreographer," the system will conclude that the label does apply. However, if an erroneous conclusion was deduced from the features of just one of the positively labeled occurrences, the true picture could look different.

Selecting only the label that is attributed in the highest ratio of occurrences may alleviate the burden of classification. For instance, the classifiers of the *Etcetera* corpus attribute three positive labels to the Flemish scenographer Johan Daenen: that of scenographer, performer and director. When the attributions are examined more closely, we find that, on a total of twenty-three occurrences (and, hence, as many twenty-word windows), Daenen is labeled as a 'scenographer' in 82.6 percent of the cases, but as a 'performer' and 'director' in only 56.5 and 52.1 percent of the cases, respectively. Thus, the system can make incorrect guesses, but it can be fine-tuned to evaluate its errors and compensate for them.

Evidently, this correction works best for artists that are mentioned frequently. For rarely mentioned names, the system simply does not have enough information to make accurate judgments. For instance, French writer Simone de Beauvoir is incorrectly labeled a 'scenographer,' while 'theorist' would be the obvious choice. Examining her case more closely, there appear to be only six snippets of text available for automatic analysis, of which four were labeled as possibly belonging to the scenographer label, and only two as belonging to theorist.

In the existing web interface for browsing and searching the *Etcetera* collection, it would obviously look very awkward if a sidebar holding metadata about the document designated Simone de Beauvoir as a scenographer. If a researcher finds that a research tool offers *some* unreliable information, the whole of the collection is affected. Hence, the user might also become skeptical of the most basic metadata, such as the name of the author or the number of the issue. This is obviously an undesirable side-effect, and it strongly cautions against the inclusion of semi-automatically integrated metadata into a digital research tool. One thinks of the controversies that plagued *Google Books* in 2009 because of widespread errors in its automatically generated metadata, such as attributing a book on the *Mosaic* internet browser to Sigmund Freud, or classifying a catalogue of copyright entries from the Library of Congress under the category 'Drama.'[16]

However, if the insertion of this kind of metadata is handled more carefully, and thoroughly sampled and tested before publication, it may make a very valuable addition to the corpus. In that case, it might be desirable to limit the automatic attribution of labels to names that are mentioned in at least ten or twenty text snippets. A different avenue for further research leads through larger sets of training data. Instead of depending on a manually labeled data set of just 458 items, a much larger database could be assembled, or an existing database could be used. In the specific case of *Etcetera*, which mainly deals with Flemish performing arts history from the 1980s and after, a possible candidate would be the data set developed by the Vlaams Theaterinstituut ('Flemish Institute for the Performing Arts').

Conclusion

In this essay, I have observed how theatrical documents, in particular reviews, can be seen as a particular form of theatrical heritage. Becoming accessible in wholly new ways through large-scale digital collections, the theater scholar's interaction with these documents has changed profoundly. I have suggested an important new

[16] Nunberg 2009.

use of digital collections, namely, text-mining as an enrichment of our understanding of theater history. The totality of all theater reviews ever published could be said to hold a virtual record of modern theater history. Through the application of techniques such as named entity recognition and information extraction to large-scale digital collections of such documents, it is indeed possible to build a database of historical theatrical events. To examine the practical aspects of such an undertaking, I presented a case study in which an existing corpus was semi-automatically enriched with semantic information. More particularly, I applied supervised machine learning techniques in order to label the extracted names with semantic roles, namely, the role each name plays in the artistic process. To conclude, I surveyed a number of future possibilities and dangers inherent in this approach.