According to the Stanford literary historian Franco Moretti, the increasing and even overwhelming availability of digital texts signifies a sea change for the scholarly study of literature and theatre. Before, it was a practical necessity to restrict a research project to at most a few dozen novels or theatre plays. Today's huge collections of digital texts, on the contrary, lend themselves to automated analysis using language technology.

What words is Shakespeare most inclined to use next to 'love'? Is the vocabulary of Macbeth closer to that of Othello, or rather that of Richard II? What characters favour conditional sentences in their speeches? What other Elizabethan authors was Shakespeare most closely related to in terms of language use?

Moretti dubbed this kind of questioning 'distant reading', implying a panoramic view of literary text that was before unfeasible. Still, in literary history and theatre studies there is still a strong focus on singular creators and exceptional works. The case study is the dominant methodological form in the majority of academic papers. But what is the significance of the exception, if we measure it against the broad mass of all plays that have been performed in a certain period? As Moretti writes, he came to this realization after the observation that the English-language canon for the nineteenth century might very well comprise around 200 novels, but that this number pales in comparison to the twenty or maybe even 30,000 novels that have been published in total during this period (Moretti 2005: 3–4).

If we approach the history of theatre not as a small canonical selection of authors, but as the total mass of theatrical events that have taken place, we will also need new tools to grasp, analyse and visualize this topic. The title of Moretti's recent pamphlet, Graphs, Maps, Trees: Abstract models for literary history (2005), suggests in what direction to search for these instruments. In the case that is under discussion here, I will deal with 'mapping' – as literally as possible – a landscape of performative texts. At 'Camillo 2.0', I applied Moretti's approach to a collection of documents from Dutch and Flemish post-war theatre. My aim was to demonstrate how map-inspired visualizations can produce a digital 'theatre of theatre memory', allowing the researcher to traverse and overview the landscape of texts in new ways.

**Mapping a Landscape of Texts**

**THOMAS CROMBEZ**

According to the Stanford literary historian Franco Moretti, the increasing and even overwhelming availability of digital texts signifies a sea change for the scholarly study of literature and theatre. Before, it was a practical necessity to restrict a research project to at most a few dozen novels or theatre plays. Today’s huge collections of digital texts, on the contrary, lend themselves to automated analysis using language technology.

What words is Shakespeare most inclined to use next to ‘love’? Is the vocabulary of Macbeth closer to that of Othello, or rather that of Richard II? What characters favour conditional sentences in their speeches? What other Elizabethan authors was Shakespeare most closely related to in terms of language use?

Moretti dubbed this kind of questioning ‘distant reading’, implying a panoramic view of literary text that was before unfeasible. Still, in literary history and theatre studies there is still a strong focus on singular creators and exceptional works. The case study is the dominant methodological form in the majority of academic papers. But what is the significance of the exception, if we measure it against the broad mass of all plays that have been performed in a certain period? As Moretti writes, he came to this realization after the observation that the English-language canon for the nineteenth century might very well comprise around 200 novels, but that this number pales in comparison to the twenty or maybe even 30,000 novels that have been published in total during this period (Moretti 2005: 3–4).

If we approach the history of theatre not as a small canonical selection of authors, but as the total mass of theatrical events that have taken place, we will also need new tools to grasp, analyse and visualize this topic. The title of Moretti’s recent pamphlet, *Graphs, Maps, Trees: Abstract models for literary history* (2005), suggests in what direction to search for these instruments. In the case that is under discussion here, I will deal with ‘mapping’ – as literally as possible – a landscape of performative texts. At ‘Camillo 2.0’, I applied Moretti’s approach to a collection of documents from Dutch and Flemish post-war theatre. My aim was to demonstrate how map-inspired visualizations can produce a digital ‘theatre of theatre memory’, allowing the researcher to traverse and overview the landscape of texts in new ways.
of documents, from many different artistic disciplines, such as happening scripts from the visual arts, and poems that were written for literary performances. The corpus that was digitized for this project currently comprises forty-five documents by six distinct authors.

**INTERTEXTUAL LANDSCAPES**

A geographical landscape can be mapped because geographical distances can be measured, and these measures can be translated into a visual representation. But how can you measure text, even a single sentence, in a non-trivial way? And how can you compare the measurements of different documents and plot them on a map? Mapping texts means comparing texts on a numerical scale, or to map them in one or more mathematical dimensions. What, then, is an adequate numerical model for text?

The dominant approach in natural language processing is to describe the document as a list of frequencies of words. I should stress that this is a very restricted way of looking at text. It is a purely lexical description, that is, in terms of its vocabulary. This means that much linguistic information is abstracted away: sentence length, sentence construction, word order, document structure, etc. All syntactic and structural aspects of the document are left out.

The model is known as the ‘bag of words’ approach. Despite its simplicity, the model has proven very successful for computational applications (Manning and Schütze 1999).

Using this model, each document is represented as a word vector. In total, there are about 29,000 distinct word forms in the whole corpus, and each document is described simply as a list of how many times each of those words occurs. This results in a numerical representation of a text document, which can be used for mutual comparisons of documents.

In two-dimensional space, each point has two coordinates – for instance, latitude and longitude. For determining the distance between two points, we simply need to apply the theorem of Pythagoras, and the result will be the Euclidean distance between two points.

The concept can easily be generalized to points in three-dimensional space and in n-dimensional space. In the case of the landscape of texts, we do not have geographical space, but ‘word space’, language considered as a multi-dimensional space, where each word is a ‘dimension’ of a document. For each of the 29,000 words occurring in the total corpus, each document has a frequency value. (Most of those are zero, but for a document not to feature a certain word is also significant information.) This means that it is possible to compute a numerical distance between two ‘points’, using their 29,000 word frequency values as ‘coordinates’.

Applied to the collection of forty-five performative texts, we can thus arrive at the intertextual distance between every pair of documents (and there are 990 of those pairs). Intertextual distance expresses how far apart, or how close, two documents are in terms of their vocabulary. In other words, it states how much or how little they overlap lexically.

The only challenge that remains is how to adequately analyse a set of 990 ‘relations’ (expressed simply as numbers) between forty-five documents. This is where techniques for mapping enter the picture, in this case network visualization.

**THE NETWORK AS A LANDSCAPE**

If we model each text as a node in the network, we can model the intertextual distances as connections between them. And the distance between those nodes can simply be equal to the distance measure which has been computed before. The closer two documents are in terms of their mutual overlap, the lower the textual distance will be, and the closer they will appear in the visualization.

In this example, I have visualized the similarity relation between one central document (a text by Hugo Claus from 1960) and twenty other corpus documents. These distances range from very closely related (scores of 3.9 or 4.0) to very distantly related (scores of 27 or 35).
On this visualization, the closeness of the relation is visualized in three different ways:

- Documents that are closely related are situated closer to the central node.
- Documents that are closely related are shown with bigger nodes, and the size of the node is proportional to the closeness.
- Documents that have a distance measure below the average are shown in blue, while the other documents are shown in red.

The result is a visualization of ‘two circles’: a first circle of closely related texts in blue, and a second circle of distantly related texts in red.

Applying this technique to the whole corpus, it is unfeasible to visualize all forty-five nodes and all 990 relations between them, since that would result in an unreadable visual mess. Therefore, I chose to visualize only the closest three relations of each of the forty-five texts.

The diagram below shows snapshots of twenty stages in the generation of the network visualization. The spring layout of the graph ‘pushes’ the nodes to the outside, while the connections try to keep them together. The result is a form of automatic clustering – some documents will stick together because there are many close relations between them; others will drift further apart because there are too few interconnections.

Eventually, the layout of the graph will reach a point of equilibrium. Both dense and sparse clusters emerge. What we learn from this visualization is the following.

First, the most visible clustering is that by authorship – the coloured nodes are forming quite neat clusters. Still, the different authors show different kind of clusterings: the texts by Tone Brulin (red), for example, are much closer together than those by Hugo Claus (bright green), indicating less lexical variation within the first author’s dramatic production. This corresponds to the classic image (in Dutch drama studies) of Claus as a ‘protean’ author,
Figure 2. Twenty stages in the development of the network map, visualizing for each of forty-five performative texts the three closest relationships.
who could master many different genres and vocabularies.

A particularly important result is the second-level clustering that becomes visible. This is a clustering that corresponds to writing style, in particular the drama texts' degree of 'experimentalism'. There is one cluster of realistic drama texts focusing on Brulin and (part of) Claus and Ivo Michiels (yellow), and there is another one of more experimental, 'ecstatic' or 'ritualistic' writing that comprises Jan Decorte (bright blue) and other texts of Claus. The texts by Bert Verminnen (purple) seem to be situated in the middle between the naturalistic and the ritualistic clusters.

One particular finding is that Ivo Michiels, traditionally regarded as a strongly experimentalist author, is not in the 'experimental zone' of the diagram. This is because the experiments of Michiels are not situated on the level of his choice of words (his vocabulary is intentionally plain and restricted) but rather on the level of syntaxis and textual structure, featuring numerous repetitions, symmetrical dialogues and other formalist techniques.