

The Complete Works of W.F. Hermans. Using Automatic Text Comparison and XML for a Voluminous Edition

Bert Van Elsacker (bertve@gmail.com)

*Huygens Instituut
The Netherlands*

Introduction

In November 2005 the first volume of the Volledige Werken (Complete Works) of Willem Frederik Hermans appeared. This publication marked the official beginning of the largest Dutch edition project ever undertaken in the field of modern literature. There are two sides to the edition: a publication in print, aimed at a general public, and a web site, where the reader can find the editorial principles, a description of the textual history and listings of editorial emendations. The project is exceptional not only because of its size, but also because right from the beginning it has been set up as an experimental digital research project.

The initial impetus came from the need for automated text comparison. An academic edition cannot do without careful comparison of the different versions of a text. In the case of Hermans, about one third of all print editions has been revised, which brings the volume of research material to about 50,000 pages. This implied manual collation was not feasible. We will give an overview of the procedure to transform a large amount of print material into accurate digital text, demonstrate a system which outputs XML-TEI-encoded collation results and show an example of the possible uses of these documents.

The edition

Willem Frederik Hermans is widely regarded as the most important Dutch author of the second half of the twentieth century. In addition to novels, Hermans (1921-1995) wrote short stories, plays, poetry and essays; he also translated several texts, including Ludwig Wittgenstein's *Tractatus Logico-Philosophicus*. His work is receiving more and more international attention; over the past few years several novels have been translated into German, French and English. Beyond *Sleep* (*Nooit meer slapen*) was hailed in the English press as a

forgotten masterpiece of post-war European literature, or in the words of the *Times* 'a welcome if belated introduction to an original and challenging voice in modern European literature'. Hermans was a very prolific author. The edition will consist of a total of 24 volumes, each with an average of about 800 pages. The Huygens Institute (a research institute within the Royal Netherlands Academy of Arts and Sciences) is responsible for the preparation of a reliable edition text in accordance to scholarly standards. For each work, the last version Hermans authorised is the point of departure for the critical edition. The editors review this version in the light of the results of archival research and of a comparison with selected other versions of the text.

Automatic text comparison

Traditionally the meticulous comparison of selected versions of a text has been one of the most important tasks of an editor. Until a few decades ago manual collation – which is extremely time-consuming – was the only possibility. However, in computer science theoretical and practical research into automated text comparison has been taking place since the 1970s. In computer science, the general task of "text comparison" has to be expressed as a formal procedure in order to make the development of an algorithm possible. More precisely, "text comparison" has been understood as a procedure which results in a list of changes between two versions of a text. The application of all changes to version A transforms version A into version B. The algorithm should try to keep the list as short as possible. This approach has led to a basic algorithm by Eugene Myers, on which some variations have been developed, and various implementations, of which the Unix/Linux tool 'diff' is the most widely used. Source code implementing this algorithm is readily available for all major programming languages.

Somewhat apart from these developments in computer science, in the world of scholarly editing there have also been initiatives to use computers for text comparison and, by extension, for the production of editions. The best-known examples are Peter Robinson's program *Collate* and Wilhelm Ott's *TUSTEP*. *Collate* is only available for the Macintosh Classic platform, an operating system which has now been replaced by OS X. The program is particularly suitable for older texts which have been divided into relatively short passages beforehand and in which there are not too many long or complex variants. The algorithm used to extract variants remains undocumented. Mr. Robinson has announced a successor to *Collate* called *EDITION*. *TUSTEP* is actually a comprehensive environment for textual research and the production of editions. It is a rather complex instrument to work with and is more something like a programming language in itself. Another impediment is the absence of a graphical user interface (GUI). Recently, another interesting

application, called Juxta, has become available. Unfortunately, for now the program lacks the option to export the collation results. According to the project web site, this will change with the next release later this year.

On the one hand automatic text comparison has great advantages: the comparison is based on a formally defined algorithm, free of errors and in principle reproducible by others. Moreover, the use of computers saves a huge amount of time, which may be of crucial importance as often resources are lacking to collate texts manually. On the other hand, there is no ready-made solution for the average user; whichever option is chosen, some extra training of the prospective user is necessary, and experience with scripting will probably come in handy. Of course this learning process also takes time and energy. For small projects the effort may outweigh the advantages, but considering the size of the Hermans project, in this case the investment did seem worthwhile. To date we have made extensive use of Collate and software tools such as diff.

Automatic collation requires accurate digital sources. For a scholarly edition, OCR-results aren't good enough. A checking system is necessary. Moreover, some presentational markup like quotes, the use of bold, italic etc., and expressive white space has to be captured as semantic markup, while other presentational features (page width, fonts used...) have no importance, in any case for the Hermans edition, and can be quietly disregarded. So a phase of checking and automatic encoding before the actual collation is required. We will discuss this operation in more detail in our poster presentation.

XML-encoded collation results

The result of these preparations are reliable and detailed overviews of all the differences between sometimes a dozen versions of a text, encoded as a base text with in-line apparatus conforming to the TEI Guidelines. This means that a complex text history can be examined down to the tiniest detail in a manageable way. The use of XML enables the editor to observe patterns in variants, to categorize findings, and to examine them in greater detail. Due to systematic encoding of the material the accumulated data can be searched accurately and checked for correlations, and working hypotheses can be continually tested (for example by using XML search languages such as XPath and XQuery) and if necessary modified.

Currently, we are working on ways to present the multiple versions of the text and the conclusions reached in research in a dynamic way, partly on the basis of the digital research documentation and the findings of the analysis of De tranen der acacia's (a major novel which appeared in the first volume). During the poster session, we will display the short story 'Paranoia'. The digital presentation of this story contains the

reading text of the Hermans edition and all the text versions which were of importance for the editorial research. We intend to place a short introductory section before the full-text documentation in which some important revisions in the short story are discussed, as a reader's guide. For example, in the version of this story in the first publication in book form in 1953 Hermans puts more emphasis on the theme of the housing shortage, which was a key concern in the Netherlands after the war, especially in Amsterdam, where the story is set. Due to adjustments in substance and narrative, the events in the book version of 'Paranoia' are described more from the perspective of the character Cleever than in the magazine publication of 1948, a significant revision in a story about someone who suffers from persecution mania.

In the digital publication we will integrate observations and analyses into the texts to which they refer, as a form of empirical evidence. Relevant text passages will therefore be tagged in the online presentation so that they can be seen separately and in context. We are also examining other presentation options. There are analyses conceivable, such as a narratological study of narrative structure, which in the form of hypertext can serve as a point of access or a guide to Hermans's work. Ideally, in future digital text presentations text and research will constitute an integrated collection of data which can constantly be consulted, modified and expanded. Or as Hermans himself once put it: '...a collection, an enormous accumulation of movements and ideas.'

Bibliography

Juxta. <<http://www.patacriticism.org/juxta/>>

Myers, Eugene W. "An O(ND) Difference Algorithm and Its Variations." *Algorithmica* 1.2 (1986): 251-266.

Ott, Wilhelm. "Strategies and Tools for Textual Scholarship: The Tübingen System of Text Processing Programs (TUSTEP)." *Literary & Linguistic Computing* 15.1 (2000): 93-108.

Robinson, Peter. *Collate 2: A User Guide*. Oxford: The Computers and Variant Texts Project, 1994.