Word Spotting as a Tool for Scribal Attribution

Lasse Mårtensson [1], Anders Hast [2], Alicia Fornes [3]
[1] Högskolan i Gävle, Sweden; [2] Uppsala Universitet, Sweden; [3] Universitat Autònoma de Barcelona, Spain;
laeman@hig.se

Word Spotting is a set of methods for localizing word forms in handwritten text. The project group behind the current abstract has previously used it on medieval Swedish manuscripts, namely on Cod. Ups C 64 (Latin) and Cod. Ups. C 61 (Old Swedish), see Wahlberg et al. (2011) and Wahlberg et al. (2014). The most common usage for Word Spotting is to extract words for different purposes, for instance for linguistic investigations. From a technical perspective, there are several different variants of Word Spotting, but in most cases the searching process is built up on a template of the word form in question being chosen, and then the computer identifies graph sequences in the manuscript, charter etc. that are similar to the template. For further details on the technical aspects of the method, see below.

In the present investigation, the Word Spotting method is used for another purpose, namely scribal attribution, i.e. identifying individual scribes. Our material is the medieval Swedish charter corpus in its entirety, as far as they have been photographed (more than 10 000 charters). These are preserved at Svenskt diplomatarium, Riksarkivet. As stated above, the basic concept of the Word Spotting method is that a word template is chosen as a point of reference, from which the other similar word forms are identified. From a linguistic perspective, the template consists in a graph sequence, as such unique and produced by a certain scribe at a certain time. This means that the template contains some characteristics of the scribe that produced it. For our purpose, the template is not used for identifying all the word forms in the corpus that the template represents, but for identifying the instances when the word forms (and individual letters; see below) have been executed in a way similar to the template.

For the purpose of scribal attribution, not only graph sequences (in this case word forms) are of interest, but also individual graphs (letters). The shape of letters has for a long time been considered as a key issue for scribal attribution in the palaeographic research. One could mention Per-Axel Wiktorsson's work in four volumes, Sveriges medeltida skrivare (2015), where Wiktorsson identifies the scribes mainly on the basis of the shape of seven letters: 'g', 'w-', 'æ', 'ø', 'y', ' n', 'k' och 'h' (p. 27). We have therefore focused on the identification of specific letters, and especially those consisting of several components, with a more complicated ductus, more specifically those used by Wiktorsson. These are, of course, more likely to show individual traits than more simple formations such as 'i', 'o' etc. In our investigations this far, we have made searches for 'g', 'æ' and 'k', and we will continue with the other ones listed by Wiktorsson.

From a technical perspective, the search for individual letters poses a greater challenge than the search for sequences, as the number of measuring points for the former is much smaller. Thus, a great deal of time has been put into optimizing the technical aspects of the method. The current state of the art in HTR (Handwritten Text Recognition; Llados et al. 2012) can be divided into at least two categories: 1) Segmentation techniques (Rath et al. 2007) need to segment the documents into text lines or even into words. Therefore, the performance of these techniques highly depends on the accuracy of the line or word segmentation algorithms. To this approach belong the above mentioned Wahlberg et al. (2011) and Wahlberg et al. (2014). 2) Segmentation-free approaches (Leydier et al. 2009) divide the manuscript into zones, or cells. Our approach belongs to this category and we use a so-called sliding window to match the template with the content of the window, in this case the handwritten document being investigated. The unique quality of our approach is that we can perform what has been done for a long period of time in the area known as image registration. In image registration, template images are matched to find identical images. In our case, dealing with handwritten text, this must be done in a different way, since the template and the word within the sliding window are not identical, since all graphs are unique and always displaying some incidental variation, however small. Therefore, the algorithm must be much more relaxed than in the case of ordinary image registration, i.e. allowing for variance (without losing accuracy). The current method can be used for searching for both words and graphs, and even for parts of graphs.

The fact that there are matches in the Word Spotting and the Letter Spotting process do not automatically lead to the conclusion that the letters have been produced by the same scribe. Instead the matches should be seen as suggestions, to be further evaluated by a human researcher. The matches represent graph sequences that display similarities with the template regarding the measuring points. If for instance matches are found regarding 'g' in certain documents, one would also expect matches in the same documents regarding other letters. This is, however, not always the case, and thus one must evaluate the results of the searches with great care.

One great difficulty when dealing with scribal attribution in medieval documents is the absence of ground truth. It is very rare that we know who actually held the pen in these documents, and when the scribes are known, they are in most cases known through earlier attributions. When working with new methods for scribal attribution, it is not satisfactory to rely on previous attributions only. If one would use previous attributions to evaluate the methods, one would risk going in circle, forming the new methods on the previous work. For that purpose, we have established a set of charters where the scribes have been identified on external evidence, i.e. not through attributions on palaeographic grounds etc. Most important are the charters containing a notice from a recording clerk, stating that this person has written the document in his own hand (see Wiktorsson 2015: 28). These charters function as our point of reference in the searches in the corpus.

This investigation is a part of an ongoing project, called "New Eyes on the Scribes of medieval Sweden" (Riksbankens jubileumsfond). The aim of this project is to investigate and map the characteristics of the script and the scribes in the medieval Swedish charters. Within this project, we use several methods, each aiming at measuring certain features of the script (see e.g. Mårtensson et al. 2015). Hence, the current Word Spotting investigation should not be seen as one isolated attempt at solving the issue of scribal attribution, but as a part of a large scale mapping of script features. The purpose of this project is not to find one single method that will work as an automatic tool for scribal attribution. It is through the collected evidence of several methods for measuring script features that a new mapping of the medieval scribes will be achieved.

Bibliography

Leydier, Y., A. Ouji, F. LeBourgeois, and H. Emptoz (2009). Towards an omnilingual word retrieval system for ancient manuscripts. In: Pattern Recognition, vol. 42, no. 9, pp. 2089–2105, 2009.

Llados, J., M. Rusinol, A. Fornes, D. Fernandez, and A. Dutta (2012). On the influence of word representations for handwritten word spotting in historical documents. In: International Journal of Pattern Recognition and Artificial Intelligence, vol. 26, no. 05, 2012.

Mårtensson, L., F. Wahlberg and A. Brun (2015). Digital Palaeography and the Old Swedish Script. The Quill Feature Method as a Tool for Scribal Attribution. In: Arkiv för nordisk filologi 130/2015.

Rath, T., and R. Manmatha (2007). Word spotting for historical documents. In: IJDAR, pp. 139–152, 2007.

Wahlberg, F., M. Dahllöf, L. Mårtensson and A. Brun (2011). Data Mining Medieval Documents by Word Spotting. In: Proceedings of the 2011 Workshop on Historical Document Imaging and Processing.

Wahlberg, F., M. Dahllöf, L. Mårtensson and A. Brun (2014). Spotting Words in Medieval Manuscripts. In: Studia Neophilologica 86/2014.

Wiktorsson, P.-A. (2015). Skrivare i det medeltida Sverige. Vol. 1. Skara: Skara stifthistoriska sällskap.