‘Peering into darkness’: the uses and usefulness of language technology to the Gender & Work project

Maria Ågren and Jonas Lindström
Department of History, Uppsala University
firstname.lastname@hist.uu.se

1 Introduction

According to a malicious view, historians are like the man who dropped his latch key in the street after a late night out. When another nocturnal walker saw him down on his knees under a street lamp, he asked him where he had dropped the key, and the first man pointed beyond the light cone, into darkness. ‘Then why are you searching here?’ the second man asked, and the first man answered ‘Because this is where there is light’.

The view is malicious as it does not take into account that, being an empirical discipline, historical scholarship has to have evidence – light. If there is no evidence, the historian cannot answer her question and needs to rephrase it so that it matches available sources. Sometimes she even has to address a completely new question. The historian is forced to look where there is light. But the problem is, of course, that vitally important questions may remain unanswered if one always changes the question, if one does not, sometimes, make the extra effort to extend area on which there is light. Sometimes, we need to construct a torch.

The Gender & Work project is an attempt to answer important questions about the ways in which women and men supported themselves in the distant past (c. 1550 to 1799) (Ågren et al., 2011). Since no systematic and reliable recordings of these matters were made at the time, the state of research was patchy and incomplete when the project started. In order to make progress, the project needed more empirical evidence, and more evidence could indeed be found in early modern sources, but at a high cost. It is very time-consuming to search handwritten documents for information. Nevertheless, thanks to a generous research grant from the Knut and Alice Wallenberg Foundation and the Swedish Research Council the project has been able to collect more than 16,000 verb-phrases describing people’s ‘time-use with the purpose of supporting oneself and one’s family’. Together with rich contextual data, the verb-phrases were stored in a specially designed relational database and now constitute the basis for more well-founded conclusions and new questions too. The total amount of searchable data is around 500,000. We have built a torch, but admittedly at a high cost.

A much smaller project is running in tandem with this big research project. The project is a close cooperation between the GaW historians and two scholars of language technology (Pettersson, Nivre), the long-term purpose being to construct a tool that can automatically identify verb-phrases of relevance to the GaW-project, i.e. verb-phrases that describe ‘work’ or, to be more precise, time-use with the purpose of supporting oneself and one’s family. Such a

1 Constructed by the Demographic Database at Umeå University.
tool would allow not only the GaW project but historians in general to find the desired type of data much more quickly than the manual search methods usually applied. The promising results of this project are described in another paper to this conference (Pettersson et al., 2014). In this paper, we will describe how we realized that a cooperative project would be of mutual interest.

2 Verbs!

At a very early stage of the GaW project, we were fortunate to have Ingrid Almqvist (then at Uppsala university, now at Stockholm university) on the advisory board. She made the shrewd remark ‘Oh, you are interested in verbs’.

The background to this remark was our presentation of the challenges involved when one tries to find out how women and men supported themselves prior to 1750/1800. In brief, the problems consist in there being no systematic data on ‘work’ from this period: there are no occupational statistics, occupational titles are few and often misleading, and the overwhelming part of all ‘work’ was unpaid and has therefore left no or a very thin paper trail. Moreover, early modern sources often pay much attention to the heads of households (usually men), to formal institutions (like guilds) and to payment of taxes. Consequently, they shield activities by women and children, activities that took place ‘in the interstices’ between households (sometimes of a criminal character) and activities that were of no interest from the point of view of fiscal authorities.

However, a pioneering study of southwestern Germany (Württenberg) by Sheilagh Ogilvie (2003) had shown that a way around these problems could be to collect information on activities/time-use from court records; often, such information was given in passing when a witness described what s/he had been doing when s/he observed a crime taking place. Ogilvie collected a little less than 3,000 observations with this method.

Inspired by Ogilvie, the GaW project set out to do something similar although on a larger scale. We were interested in observations that took the form of, for instance, ‘milk cow’, ‘mow grass’, ‘fish herring’, ‘sell herring’, ‘knit mittens’, etc. But we had not realized that the desired data could be described as verb-phrases, typically verb + complement. This was where Ingrid Almqvist’s comment was so useful. She identified the point where the historians’ need for more and better evidence and the language technologists’ interest in new scholarly problems intersected. Ingrid Almqvist suggested a cooperative project with Joakim Nivre and Eva Pettersson. Her remark also inspired us to call our method ‘the verb-oriented method’.

3 A pilot project followed by a full-scale project

A pilot project was carried out in 2010. At this stage, the problem to be solved was two-fold. First, was it possible to identify the verbs in a 17th-century text with language-technological tools adapted for modern Swedish, bearing in mind the differences in spelling, vocabulary and syntax? Second, was it possible to digitize the pilot source (a 17th-century court record from Uppland), using OCR, without a drastic decline in text quality?

The identification of verbs, based on a combination of two methods, exceeded expectations: both coverage and precision reached 90 per cent (Pettersson, 2010). The OCR-scanning also gave surprisingly good results: scanning and OCR-reading took 45 minutes, while manual digitization (+ two rounds of proof reading) required 80 hours’ work. Having the text digitized with OCR reduced the time with 99 per cent, without a significant drop in quality. While manual
extraction of verbs from 100 pages with 100 per cent coverage would yield 100 observations, automatic extraction of verbs from 10,000 pages with 90 per cent coverage would yield 9,000 observations. It was easy to see the potential for reducing time and, consequently, costs.

Following this successful pilot project, a Ph.D. project was designed with more ambitious objectives. From the historians’ point of view, the most interesting aspect was whether the verb-extraction tool could ‘be taught’ to tell the difference between verb-phrases that describe ‘time-use with the purpose of supporting oneself and one’s family’ – what we in everyday language often call work – and verb-phrases that have nothing to do with work. Ongoing work on this kind of verb phrase ranking, using the manually excerpted verb-phrases in the GaW database as a key, shows promising results.

4 The future

For the GaW project, this means that we have reached a point where the verb extraction tool is useful. Analyzing a 17th or 18th-century text with the tool will allow us quickly to identify verb-phrases of probable interest. The next step is, therefore, to integrate the tool with the GaW database to allow future scholars to speed up the analysis of the texts upon first reading.

Automatic extraction of ‘work’-verbs from historical sources will not, however, make careful reading superfluous. Even if most activities of interest do take the form of a verb-phrase, and even if the program can be trained to recognize them, there are sometimes contextual dimensions that need to be taken into account. The GaW database includes a remarkable collection of verb-phrases, some of which are rather odd and can only be identified as relevant by a human reader. To give but one example: ‘to milk into ear’. This verb-phrase has been correctly classified as ‘work’: a woman poured her own breast milk into the wounded ear of another person, thus providing a form of medical care. To give another example: ‘to close the door’. Whether this phrase describes work cannot be decided once and for all. If a maid servant closed the door to the henhouse, this was probably part of her job as a servant. But if somebody closes his or her door because it is getting cold, it is hard to see that this should be classified as a work or as ‘time-use with the purpose of supporting oneself’.

These examples should not be read as arguments against the usefulness of language technology to historians. Saving us time and money by swiftly identifying the majority of verbs of interest, language technology will allow the historians to spend more time on contemplating the odd and intriguing examples.

References


