Natural Language Processing for Historical Texts

Michael Piotrowski (Leibniz Institute of European History) Morgan Claypool (Synthesis Lectures on Human Language Technologies, edited by Graeme Hirst, volume 17), 2012, ix+157 pp; paperbound, ISBN 978-1608459469.

Laurent Romary

Book Review

Natural Language Processing for Historical Texts

Michael Piotrowski
Leibniz Institute of European History

Morgan & Claypool (Synthesis Lectures on Human Language Technologies, edited by Graeme Hirst, volume 17), 2012, ix+157 pp; paperbound, ISBN 978-1608459469

Reviewed by
Laurent Romary
Inria & Humboldt University Berlin

The publication of a scholarly book is always the conjunction of an author's desire (or need) to disseminate his experience and knowledge and the interest or expectations of a potential community of readers to gain benefit from the publication itself. Michael Piotrowski has indeed managed to optimise this relation by bringing to the public a compendium of information, which I think has been heavily awaited by many scholars having to deal with corpora of historical texts. The book covers most topics related to the acquisition, encoding and annotation of historical textual data, seen from the point of view of their linguistic content. As such, it does not address issues related, for instance, to scholarly editions of these texts, but conveys a wealth of information on the various aspects where recent developments in language technology may help digital humanities projects to be aware of the current state of the art in the field.

Still, the book is not an encyclopedic description of such technologies. It is based on the experience acquired by the author within the corpus development projects he has been involved in, and reflects in particular the specific topics on which he has made more in-depth explorations. It is thus written more as a series of return on experience than a systematic resource to which one would want to return after its initial reading.

The book is organized as a series of 9 short chapters, which I describe below.

In the first two (very short) chapters, the author presents the general scope of the book and provides an overview of the reasons why NLP has such an entrenched position in digital humanities at large and the study of historical text in particular. Citing several prominent projects and corpus initiatives that have taken place in the last few decades, Piotrowski defends the thesis, which I share, that a deep understanding of textual documents requires some basic knowledge of language processing methods and techniques. Chapter 2 in particular ("NLP and digital humanities") could be read as an autonomous position paper, which, independently of the following chapters, presents the current landscape of infrastructural initiatives and scholarly projects that shape this convergence between the two fields.

Chapter 3 ("Spelling in historical texts"; pp. 11–23) describes the various issues related to spelling variations in historical text. It shows how difficult it may be to deal with both diachronic (e.g. in comparison to modern standardised spellings) and synchronic variations (degree of stabilisation of historical spellings), especially in the context of the uncertainty brought about by the transcription process itself. This is particularly true for historical manuscripts and Piotrowski goes deeply into this, showing some concrete examples of the kind of hurdles that a scholar may fall into. This is the kind of short introduction I would recommend for anyone, in particular students, wanting to gain a first understanding in the domain of historical spelling.

Chapter 4 is the longest chapter in the book "Acquiring historical texts"; pp. 25–52) and covers various aspects of the digitization workflow that needs to be set up for creating a corpus of historical texts. The chapter is quite difficult to to read as one single unit because of its intrinsic heterogeneity. Indeed, it covers quite a wide range of topics: presentation of existing digitisation projects worldwide, technical issues related to scanning, comparison of various optical character recognition systems for various types of scripts, the potential role of lexical resources, crowd-sourcing for OCR post-processing, manual or semi-automatic keying. Getting an overview of the various topics is even more difficult because of the way the author has followed his own personal experience, and alternates between general considerations and in-depth presentations of concrete results. Pages 34–40 for instance is one single subsection on the comparison of OCR outputs that goes into so much detail that it breaks out the continuity of the argument although in itself this subsection could be really interesting for a specialized reader. As we shall see in the conclusion, this chapter illustrates that the content of this book would benefit from being published on a more modern and opened setting.

Data representation aspects are covered in chapter 5 ("Text encoding and annotation schemes"; pp. 53–68), which tackles two specific issues, namely character and document encoding. On these two, the author presents what could be considered best practices. For character encoding, the book rightly focuses on the advantages that the move towards ISO 10646/Unicode has brought to the community. The corresponding subsection actually covers three different aspects: it first makes an extensive presentation of the history of character encoding standards (from ASCII/ISO 646 to Unicode/ISO 10646), provides insights into the current coverage and encoding principles (e.g. UTF-8 vs. UTF-16) of ISO 10646, and finally focuses on the specific difficulties occurring in historical texts both from the point of view of legacy ASCII based transcription languages and the management of characters that are not present in Unicode. Although well documented, these three topics should have been more clearly separated so that readers interested in one or the other could directly refer to it. This is a typical case where, given the great expertise of the author in the subject, I could imagine the corresponding texts being published online as separate entries in a blog. The second half of the chapter focuses on the role of the TEI guidelines for the transcription and encoding of historical text. It covers the various representation levels that may be concerned (metadata, text structure, surface annotation) and insist on the current difficulty of linking current NLP tools to TEI encoded documents. While this is indeed still an issue in general, it might have been interesting to refer to standards (ISO 24611– MAF) and initiatives (Textgrid core encoding at the token level; the TXM platform for text mining) that have started to provide concrete sustainable answers to the issue.

The following chapter ("Handling spelling variations"; pp. 69–84), provides a series of short studies describing possible methods for dealing with OCR errors or spelling variation as described in chapter 3. Independently from the fact that I find it strange to see the two chapters quite far from one another, the present one distinguishes itself with its profound heterogeneity. While several sections do have the most appropriate level of details and topicality for historical texts (in particular those on canonicalization), some sections seem to be completely off-topic (section 6.2, "Edit Distance", describes what I would consider as background knowledge for such a book). It is all the more disappointing that the author shows here a very high level of expertise and as in the case of chapter 3, I would strongly recommend the reading of the relevant sections to newcomers in the field.

In contrast with the previous chapter, chapter 7 ("NLP tools for historical languages"; pp. 85–100) is more coherent and focused. It mainly addresses the morpho-

syntactic analysis of historical text and presents, through concrete deployment scenarios, possible methods to constrain the appropriate parsers, in a context where hardly any existing tools can be simply re-used. The chapter is very well documented and refers to most of the relevant initiatives in the domain of morphology for historical text, at least on the European scene. This focus may also be misleading since recent work on named entity recognition on historical texts are not at all mentioned and are probably, to my view, one of the most promising direction for enhanced digital scholarship.

The last chapter ("Historical corpora"; pp. 101–116) is a compendium, sorted by language, of the major historical corpora available worldwide. It shows the dynamic that currently exists in the community and is an essential background resource to both understanding who is active in maintaining historical corpora and discerning the most relevant resources. The chapter as a whole provides an interesting "historical" perspective on the progress made by most text-based projects in using the TEI guidelines as their reference standard. It seems quite difficult now to imagine an initiative which would not take TEI for granted, and would not build inside the TEI framework. On another issue, namely copyright, Piotrowski also provides an interesting analysis on the difficulty of re-using old editions which have been recently re-edited on paper, and thus fall into some publisher's copyright restrictions. The conclusion could have been a little tougher here though, and probably should have recommended putting a hold on any paper publication of historical sources by a private publisher, unless it is guaranteed that the electronic material can be used freely, under an appropriate open license.

As a whole, the book leaves the reader with a mixed feeling of enthusiasm and disappointment. Enthusiasm, because the content is so rich that it should serve as background reference (and indeed be quoted) for any further work on the creation, management and curation of historical corpora. Still, I cannot help thinking that the editorial setting as a book is not the most appropriate setting for such content. The variety of topics that are addressed as well as the heterogeneous levels of details provided through the different chapters would benefit from a more fragmented treatment. Indeed, this would be the perfect content for a series of blog entries (for instance in a scholarly blog such as those on the hypotheses.org platform) which in turn would allow an interested reader to discover exactly the topics he wants information about and cite the corresponding entries. With the bibliography in Zotero and relevant pointers to the corresponding online corpora or tools, I could imagine the resulting content soon becoming one of the most cited online resources. I am sure the author would gain more visibility in doing so than having the material hidden on a library shelf or behind a paywall. Not knowing the exact copyright transfer agreement associated with the book, I cannot judge if it is too late for the author to think in these terms, but this could be a lesson for scholars who are now planning to write such an introductory publication. Is the book still the best medium?

This book review was edited by Pierre Isabelle

Laurent Romary is Directeur de Recherche at Inria, France, and guest scientist at the Humboldt University in Berlin. He has been involved for many years in language resource modeling activities and in particular in standardization initiatives in the TEI consortium and ISO committee TC 37/SC 4 (language resource management). He is the director of the European DARIAH digital infrastructure in the humanities. email: laurent.romary@inria.fr.