Modelling Scale in Historiographical Data

Florentina Armaselu
Luxembourg Centre for Contemporary and Digital History
University of Luxembourg

Introduction

The project will investigate the meaning of *scale* in historical writings, and more precisely how scale is expressed through language in historical discourse. This question draws attention to the conceptual and linguistic mechanisms at play in building historical knowledge, when the historian moves between different layers of analysis, narration or consulted sources, involving different degrees of generality. Studies in "global microhistory" (Trivellato, 2011) have highlighted the idea of variable scale representation in texts that integrate micro and global history. Nonetheless, the question of how this phenomenon of scale is actually expressed through language in historians' discourse has been less studied so far.

A small historiographical corpus, in which variations of scale are clearly present, will serve to develop the digital approach/tools/methodology. Depending on the findings of the project, an extension of the research to other types of corpora is envisaged.

The proposal presents the intended workflow for a scale analysis prototype and the first experimental results using topic modelling.

Workflow

The historiographical corpus includes a series of books that mingle micro and macro scales of historical analysis (Brook, 2009; Rothschild, 2013; Sparks, 2004; Stein, 2008; Wills, 2001). Each document is cut in chapters (or parts), considered as units of analysis to be examined via a topic modelling tool. The assumption is that, for this type of documents, the text inside the units can be restructured on levels of generality/specificity (or macro/micro), according to a number of topics that traverse the whole conceptual space (Figure 1).

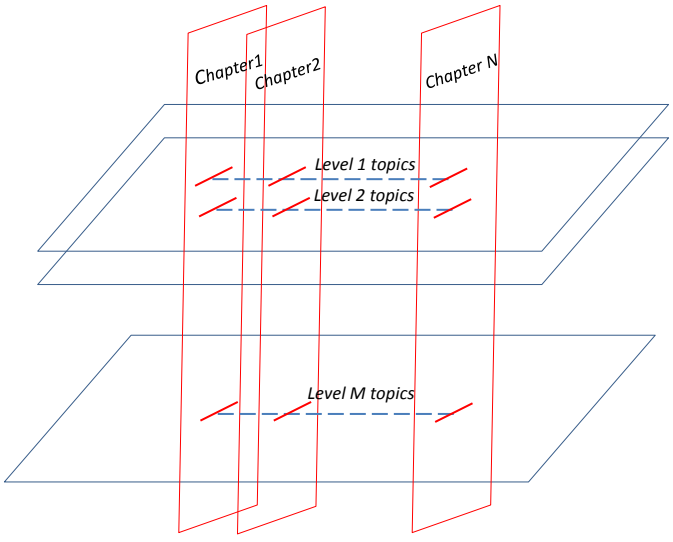

Figure 1. Restructured text by levelled topics

In a first phase, the workflow will consist in applying LDA-based techniques for topic modelling (Blei et al., 2003a, b; Blei, 2012; Rhody, 2012; Underwood, 2012; Weingart, 2012) to the selected corpus. For illustration, it is intended to provide means of exploring restructured text fragments through a zoomable layout, from global to micro perspectives (see next section). Other methods, e.g. for hierarchical or

knowledge-based topic modelling (Hofman, 1999; Hu et al. 2016) and visualisation (Smith et al., 2014; Yang et al. 2017), are also studied.

Experiments

Table 1 shows the first results of an experiment on *Vermeer's Hat* (Brook, 2009) using MALLET (McCallum, 2002) and TWiC (Topic Words in Context) (Armoza, 2017). One can observe the coverage of about 80% by the top 3 topics in each chapter and the recurrence of topics 13 and 7 in all the eight units.

Table 1. Top 3 topics for *Vermeer's Hat*, by chapters

Document (chapter)	Topic	Proportion	Topic	Proportion	Topic	Proportion	Total top 3 topics by chapter
1_TheViewFromDelft_0.txt	**13**	0.418965239	**14**	0.294147794	**7**	0.145840564	**0.858953597**
2_VermeersHat_0.txt	**17**	0.445442596	**7**	0.259011646	**13**	0.1249389	**0.829393141**
3_ADishOfFruit_0.txt	**4**	0.420070653	**7**	0.295845845	**13**	0.167718771	**0.883635269**
4_GeographyLessons_0.txt	**16**	0.423656784	**7**	0.306487232	**13**	0.108303131	**0.838447147**
5_SchoolForSmoking_0.txt	**8**	0.415263013	**7**	0.251512379	**13**	0.123663831	**0.790439223**
6_WeighingSilver_0.txt	**2**	0.376408245	**7**	0.332168176	**13**	0.120215764	**0.828792186**
7_Journeys_0.txt	**19**	0.378046282	**7**	0.257079274	**13**	0.205071495	**0.840197051**
8_EndingsNoManIsAnIsland_0.txt	**3**	0.379550072	**13**	0.268323093	**7**	0.217947775	**0.86582094**

Table 2. Weight and words for dominant topics in *Vermeer's Hat*

Topic	Weight	Words
7	2.4119	chinese china people back trade made european europe europeans men end time south century called make year dutch coast spanish
13	1.83863	world century vermeer dutch seventeenth painting place delft life time voc long years side east home vermeers things trade paintings
2	0.14006	silver spanish manila economy money coins gold potos god supply spain orozco parian grams coin spaniards philippines governor cargo spring
16	0.13547	foreigners macao portuguese lu las knowledge cortes xu jesuit red pan gua silver maps macanese pirates zhaolong ashore mission ships
19	0.07524	dutch ship bontekoe ships bramer painting weltevree cocchi magi black sailors korean boy men macao women story servant survivors junk
3	0.07299	donne corcuera state wen language death donnes translators trading responsibility bell treasure zhenheng governor translation bullion metaphor corcueras coin meditation
4	0.04935	porcelain ships white objects dishes lion wen dutch voc pieces amsterdam potters portuguese foreign dish taste cargo produced style li
14	0.0486	delft shanghai paintings pearl canal rotterdam schouten chamber herring schiedam oude surface kolk buildings great built influence plague cold objects
8	0.03339	tobacco smoking smoke opium beijing yang plant pipe smokers smoked women fujian tobaccos english capital native americas local things habit
17	0.02491	champlain french lake beaver native champlains arquebus huron mohawks hurons hat st montagnais chiefs lawrence allies war felt hats north

A closer examination of the words and contexts corresponding to each topic (Table 2, Figure 2) may lead to the assumption that the recurrent topics 7 and 13 correspond to higher levels in the hierarchy, i.e. accounting for larger-scale events in seventeenth century and the trade routes between Europe and China, while the other dominant topics in the chapters refer to more particular circumstances, places or people as pieces in the global picture.

Figure 3 illustrates another way of exploring the book using an editor for zoomable texts (Armaselu, 2010). The representation (created manually) shows on level 1 a general account (chapter 1) on the global cooling in the seventeenth century and gradually narrows down to the Little Ice Age in Europe, Pieter Bruegel the Elder's winter landscapes and Vermeer's *View of Delft* painted in 1660.

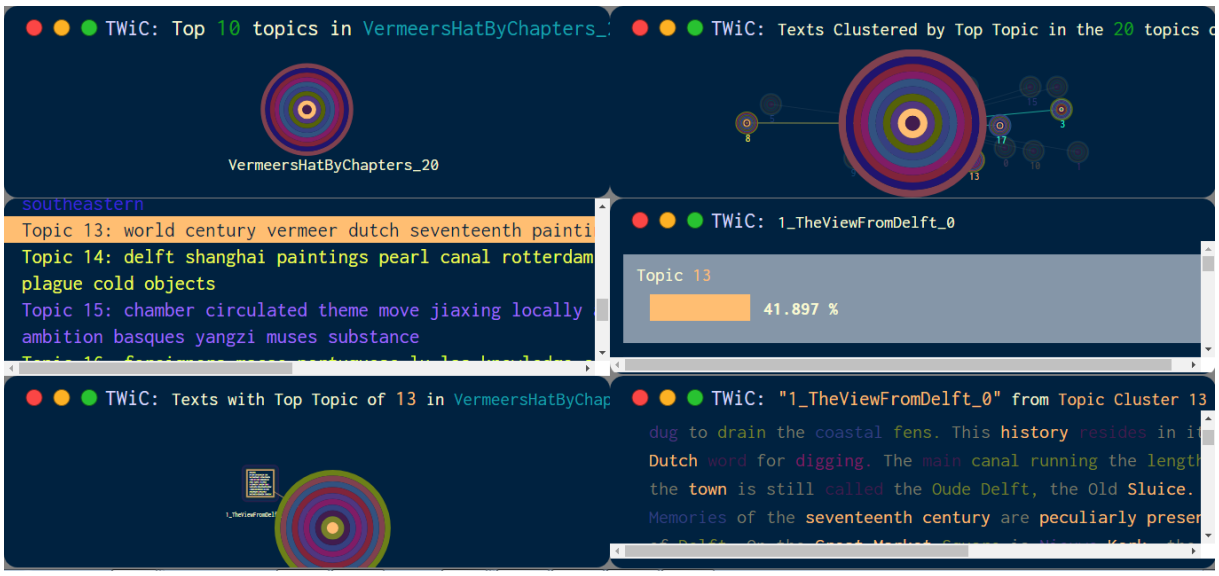

Figure 2. *Vermeer's Hat.* TWiC visualisation of topics, words, proportions and contexts

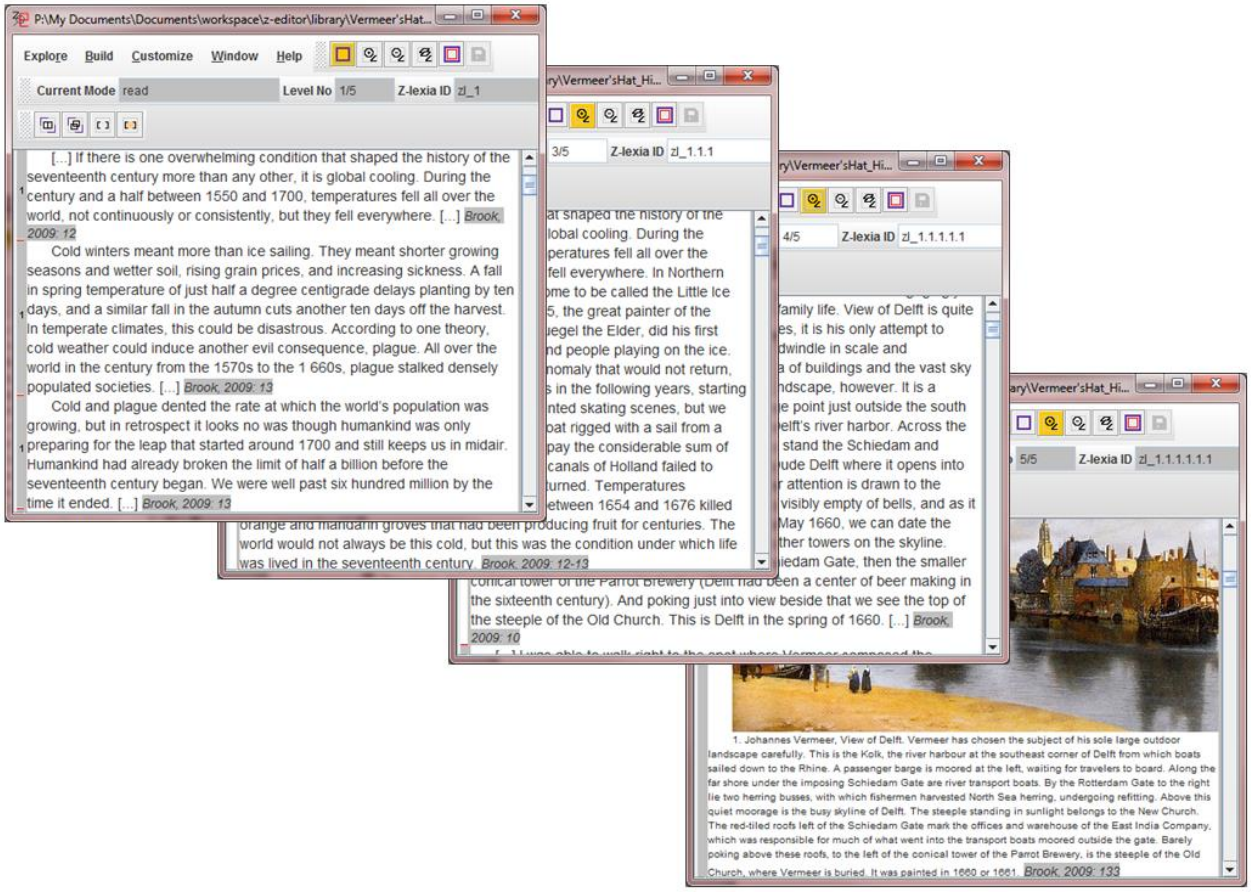

Figure 3. *Vermeer's Hat* (Brook, 2009: 12, 13, 10, 133). *Z-editor, z*oomable text layout

Conclusion and future work

The paper presents a workflow and first experiments using topic modelling to analyse scale representation in historiographical data. The project is still under development. Further work will include tests with other documents from the "global microhistory" corpus. Experiments with hierarchical topic models and other visualisation tools, as well as creating a pipeline for semi-automatic restructuration of data as zoomable texts are also envisaged.

References

Armaselu (Vasilescu) F. *Le livre sous la loupe : Nouvelles formes d'écriture électronique*, Ph. D. Thesis, Papyrus, University of Montreal Institutional Repository, https://papyrus.bib.umontreal.ca/xmlui/handle/1866/3964.

Armoza, J. TWiC, Topic Words in Context, 2017, https://github.com/jarmoza/twic.

Blei, D. M. Ng, A.Y. Jordan, M. I. "Latent Dirichlet Allocation", In *Journal of Machine Learning Research* 3 (2003) 993-1022, 2003a, http://www.jmlr.org/papers/volume3/blei03a/blei03a.pdf.

Blei, D. M. Jordan, M. I., Griffiths, T. L. Tenenbaum, J. B. "Hierarchical topic models and the nested chinese restaurant process", In Proceeding NIPS'03 Proceedings of the 16th International Conference on Neural Information Processing Systems, pages 17-24, Whistler, British Columbia, Canada — December 09 - 11, 2003, MIT Press Cambridge, MA, USA ©2003, 2003b, https://papers.nips.cc/paper/2466-hierarchical-topic-models-and-the-nested-chinese-restaurant-process.pdf.

Blei, D. M. "Topic Modeling and Digital Humanities", In *Journal of Digital Humanities (JDH)*, Vol. 2, No. 1 Winter 2012, http://journalofdigitalhumanities.org/2-1/topic-modeling-and-digital-humanities-by-david-m-blei/.

Brook, T. *Vermeer's Hat. The Seventh Century and the Dawn of the Global World*, Profile Books, 2009.

Hofmann, T. "The cluster-abstraction model: Unsupervised learning of topic hierarchies from text data". In *IJCAI*, pages 682–687, 1999, https://www.ijcai.org/Proceedings/99-2/Papers/004.pdf.

Hu, Z. Luo, G. Sachan, M. Xing, E. Nie, Z. "Grounding topic models with knowledge bases", In *Proceeding IJCAI'16*, Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence Pages 1578-1584. New York, New York, USA — July 09 - 15, 2016, AAAI Press ©2016, https://www.cs.cmu.edu/~zhitingh/data/ijcai16ground.pdf.

McCallum, A. K. "MALLET: A Machine Learning for Language Toolkit", 2002, http://mallet.cs.umass.edu.

Rhody, L.M. "Some Assembly Required: Understanding and Interpreting Topics in LDA Models of Figurative Language", August 22, 2012, http://www.lisarhody.com/some-assembly-required/.

Rothschild, E. *The Inner Life of Empires. An Eighteenth Century History*, Princeton University Press, 2011, paperback 2013.

Sparks, R. J., The Two Princes of Calabar. *An Eighteenth-Century Atlantic Odyssey*, Harvard University Press, 2004.

Stein, S. A. *Plumes. Ostrich Feathers, Jews, and a Lost World of Global Commerce*, Yale University Press, 2008.

Smith, A. Hawes, T. Myers, M. "Hierarchie: Interactive Visualization for Hierarchical Topic Models", In *Proceedings of the Workshop on Interactive Language Learning, Visualization, and Interfaces* , pages 71–78, Baltimore, Maryland, USA, June 27, 2014. c©2014 Association for Computational Linguistics, http://www.aclweb.org/anthology/W14-3111.

Trivellato, F. "Is There a Future for Italian Microhistory in the Age of Global History?", In *California Italian Studies*, 2(1), 2011, https://escholarship.org/uc/item/0z94n9hq.

Underwood, T. "Topic modeling made just simple enough", April 7, 2012, https://tedunderwood.com/2012/04/07/topic-modeling-made-just-simple-enough/.

Yang, Y. Yao Q. Qu H. "VISTopic: A visual analytics system for making sense of large document collections using hierarchical topic modelling", In *Visual Informatics* 1 (2017) 40–47.

Weingart, S. "Topic Modeling for Humanists: A Guided Tour", July 25, 2012, http://www.scottbot.net/HIAL/index.html@p=19113.html.

Wills Jr., J. E. *1688. A Global History*, W.W. Norton & Company, 2001.