Historical Models and Serial Sources

Michael Piotrowski

Journal of European Periodical Studies, 4.1 (Summer 2019)

ISSN 2506-6587

The *Journal of European Periodical Studies* is hosted by Ghent University

Website: ojs.ugent.be/jeps

To cite this article: Michael Piotrowski, 'Historical Models and Serial Sources', *Journal of European Periodical Studies*, 4.1 (Summer 2019), 8–18

Historical Models and Serial Sources

Michael Piotrowski
University of Lausanne
michael.piotrowski@unil.ch

ABSTRACT

Serial sources such as records, registers, and inventories are the 'classic' sources for quantitative history. Unstructured, narrative texts such as newspaper articles or reports were out of reach for historical analyses, both for practical reasons — availability, time needed for manual processing — and for methodological reasons — manual coding of texts is notoriously difficult and hampered by low inter-coder reliability. The recent availability of large amounts of digitized sources allows for the application of natural language processing, which has the potential to overcome these problems. However, the automatic evaluation of large amounts of texts — historical texts in particular — for historical research also brings new challenges. First of all, it requires a source criticism that goes beyond the individual source to consider the corpus as a whole. It is a well-known problem in corpus linguistics to determine the *balancedness* of a corpus, but when analyzing the content of texts rather than solely the language, determining the *meaningfulness* of a corpus is even more important. Second, automatic analyses require operationalizable descriptions of the information one is looking for. Third, automatically produced results require interpretation, in particular when — as in history — the ultimate research question is qualitative, not quantitative. This, finally, poses the question whether the insights gained could inform formal, i.e. machine-processable models, which could serve as foundation and stepping stones for further research.

KEYWORDS

Digital humanities, formal models, corpora, natural language processing

Introduction

Quantitative history has traditionally relied on serial sources such as records, registers, and inventories. Due to its roots in economic history,[1] unstructured, narrative serial sources such as newspaper articles or reports may have been traditionally less interesting, but they have also been out of reach for historical analyses, both for practical reasons — availability, time needed for manual processing — and for methodological reasons — manual coding of texts (which *is* being used in journalism studies) is notoriously difficult and hampered by low inter-coder reliability.

As more and more historical sources become available in digitized form, it is now possible to apply automatic methods from natural language processing (NLP) — text mining, information extraction, sentiment analysis, etc. — to address these problems and, at least in principle, to quickly analyze large amounts of texts with minimal human intervention. The automatic analysis of large amounts of texts has long been common in linguistics—to the point that one may say that corpus linguistics, i.e., the subfield of linguistics that is based on empirical studies of language through large corpora, has really become linguistics. The use of similar approaches in historical research, however, comes with its own set of challenges. Apart from the particular problems that historical texts pose to NLP,[2] it first of all requires a new form of source criticism that goes beyond the individual source to consider the corpus as a whole. It is well-known in corpus linguistics that *balancedness* and *representativeness* of a corpus are difficult to assess; however, determining the *meaningfulness* of a corpus for historical research is even harder. Second, automatic analyses require formal models of the phenomena of interest. Third, automatically produced results require interpretation: in historical research, research questions are ultimately qualitative, not quantitative. This, finally, poses the question whether the qualitative insights gained by quantitative analyses could themselves be formalized in order to become machine-processable and thus serve as foundation and stepping stones for further research.

This paper should be understood as a position paper: it is likely to pose more questions than it provides answers. Its main goal is to encourage a critical reflection and discussion of digital approaches towards serial publications, to pick up the title of the workshop from which it originates. The rest of this paper is structured as follows: I will first outline my understanding of digital humanities and briefly introduce formal models. I will then discuss corpora as models in linguistics and history, and finally present a (preliminary) conclusion.

Digital Humanities

The topic of this volume are *digital* approaches towards serial publications, a topic that falls under the general heading of *digital humanities* (DH), a term still waiting to gain a commonly accepted definition. More often than not, however, it simply refers to digitization of research objects in order to automate mechanical tasks and to apply quantitative methods at a larger scale — if it is not merely a fancy name for the contemporary humanities: in the 'digital age', surely the humanities must be 'digital', too. We believe, however, that the digital humanities are not just about digitization and automation, but rather about the development and use of *new methods*. In the light of the

1 Jean Marczewski, *Introduction à l'histoire quantitative: Travaux de droit, d'économie, de sociologie et des sciences politiques* (Geneva: Droz, 1965), p. 35.

2 Michael Piotrowski, *Natural Language Processing for Historical Texts, Synthesis Lectures on Human Language Technologies* (Lexington, KY: Morgan & Claypool, 2012).

digital transformation of society as a whole, the humanities are perhaps more important than ever — but they can meet their scholarly challenges and societal responsibilities only when the digitalization of the humanities is understood as an actual transformation; in particular, the humanities need to seize the opportunity to become more *agile*, to borrow a concept from software development.[3] To become more agile would mean to create explicit —ideally *computational* — models, test them rigorously, publish them early for comments, and iteratively and incrementally improve them by integrating feedback from testing and from other scholars. Computational models thus play a central role in this transformation of the humanities, and if the humanities want to tap the full potential of the computer — which lies in its use as an infinitely flexible modelling tool — they first of all need to reflect on their discipline-specific modelling challenges and practices. We therefore define *digital humanities* in the following precise fashion:

1. research on and development of means and methods for constructing formal models in the humanities (*theoretical digital humanities*), and
2. the application of these means and methods for the construction of *concrete* formal models in the humanities disciplines (*applied digital humanities*).

As a metascience for applied DH, theoretical DH — i.e., theory formation in DH — is crucial for laying the theoretical groundwork for the development of models and methods *appropriate* for the humanities — instead of using 'second-hand' methods originally developed for completely different purposes. This way, DH will in fact play an important role for the transformation of the humanities and social sciences within the larger digital transformation of society as a whole, and mean more than just 'contemporary humanities'.

This parallels the relationship between computational linguistics and corpus linguistics: the former studies the means of methods of constructing formal models in linguistics — such as formal language theory — while the latter is concerned with the construction of formal models — such as formal grammars — of *concrete* languages.[4]

Formal Models

The two definitions of digital humanities above place *formal models* front and center. Why? First of all, it is important to remember that all scientific and scholarly research constructs models of the objects of research. As Neil Gershenfeld notes, '[t]he most common misunderstanding about science is that scientists seek and find truth. They don't — they make and test models.'[5] In order to understand a complex object (phenomenon, situation, etc.), one needs to understand its parts and how they interrelate with each other. In fact, modelling is not restricted to research — we construct (mental) models all the time. However, in science and scholarship, we want to go beyond our mental models and communicate models in order to construct *shared* (i.e. intersubjective) models. Joshua M. Epstein therefore concludes:

> The choice, then, is not whether to build models; it's whether to build explicit ones. In explicit models, assumptions are laid out in detail, so we can study exactly

3 *Agile software development* describes an approach to software development that advocates adaptive planning, evolutionary development, early delivery, and continual improvement, and that encourages rapid and flexible response to change.

4 I am slightly simplifying the terminology here.

5 Neil Gershenfeld, 'Truth Is a Model', response to the thematic issue 'What Scientific Concept Would Improve Everybody's Toolkit?' *Edge* (2011), www.edge.org, no page.

what they entail. On these assumptions, this sort of thing happens. When you alter the assumptions that is what happens. By writing explicit models, you let others replicate your results.[6]

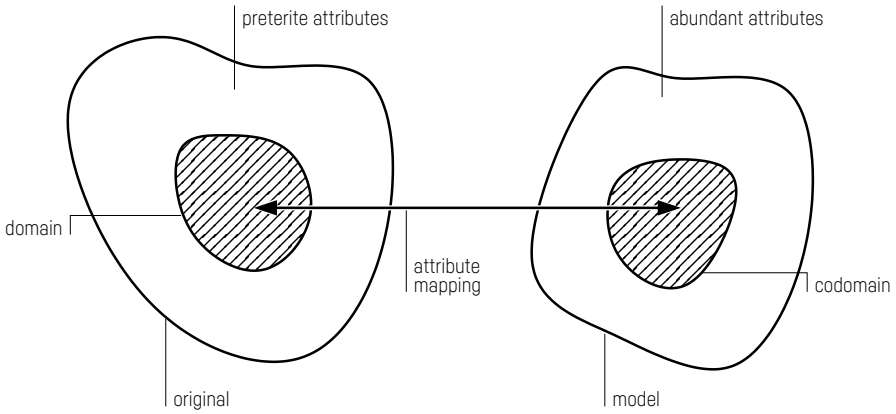

Fig. 1 Original–model mapping

What, then, do we mean by 'model'? We use the term in the sense of Herbert Stachowiak's *Allgemeine Modelltheorie* (*General Model Theory*).[7] The basic assumption is that arbitrary objects can be described as *individuals* characterized by a finite number of *attributes*.[8] Attributes can be characteristics and properties of individuals, relations between individuals, properties of properties, properties of relations, etc.[9] Modelling is then a mapping of attributes from the original to the model (see Fig. 1); in mathematical terms, it can be considered a function. Stachowiak defines models via three fundamental properties:

- Mapping property:
  Models are always models *of something*, viz mappings from, representations of natural or artificial originals, which can themselves be models.
- Reduction property:
  Models generally *do not capture all* attributes of the original they represent, but only those that the model creators and/or model users deem relevant.
- Pragmatic property:
  Models are not per se uniquely assigned to their originals. They fulfill their replacement function:
  a) for particular subjects that use the model,
  b) within particular time intervals, and
  c) restricted to particular mental or actual operations.

With respect to the mapping of attributes, three interesting cases shall be briefly mentioned, which relate to the reduction property and the pragmatic property: *preterition*, *abundance*, and *contrasting*. Preterite attributes are attributes *not mapped* from the original to the model; abundant attributes are attributes that *do not exist* in

6 Joshua M. Epstein, 'Why Model?', *Journal of Artificial Societies and Social Simulation*, 11.4 (2008), 12.
7 Herbert Stachowiak, *Allgemeine Modelltheorie* [*General Model Theory*] (Vienna: Springer, 1973).
8 Stachowiak notes that there is no intrinsic difference between individuals and attributes, individuals are merely introduced for convenience.
9 Stachowiak, p. 134.

the original. Contrasting refers to the exaggeration of certain attributes in the model, typically to highlight certain aspects of the original.

Computers allow us to create models that can be processed and manipulated by and through computational means: 'computers are essentially modelling machines, not knowledge jukeboxes'.[10] However, for a model to be 'understandable' by computers requires it to be *formally specified*. A very concise definition of 'formal' is given by Aleksej Vsevolodovič Gladkij and Igor Aleksandrovič Mel'čuk, who point out that 'the word "formal" means nothing more than "logically coherent + unambiguous + explicit"'.[11] While there are different levels of formalization, it is obvious that in digital humanities we are first and foremost interested in a level of formalization that allows our models to be processed and manipulated by computers.

Corpora as Models

Digital approaches towards serial publications in historical research are in many respects inspired by corpus-based approaches in linguistics and by applications of methods and tools from natural language processing — the engineering science based on the theoretical foundations of computational linguistics — in various fields. In this section, we therefore first look at linguistic corpora as a point of reference, before we turn to history.

Linguistics studies the structure and function of language. Since language in its totality is inaccessible for study, an empirical study of language is only possible through models. The primary model of language that has emerged over the last fifty years or so is the *corpus*. Generally speaking, a corpus in linguistics is a collection of written or spoken language material in machine-readable form, assembled according to precise criteria and for the purpose of studying one or more specific linguistic phenomena. The idea of corpora predates computers, and in particular in subfields of linguistics like dialectology, corpora have a much longer history, but machine-readability is considered a sine qua non today. Corpora can thus be considered *computational models* of language: they can be analyzed and manipulated by and through computers; this is what enables 'analyses of a different type and scope than previously possible' in linguistics,[12] apart from the fact that is obviously an essential prerequisite for their use in natural language processing. A corpus differs from a mere collection of texts by having precise criteria of construction, which are motivated by its intended purpose.[13] In short, a corpus is a model — in the sense of Stachowiak — created in order to study specific aspects of a language (or some part of it).

One of the main challenges in corpus design (= modelling) is to minimize distortion in the mapping from the original (the language) to the model (the corpus). The foundations for theoretical reflection on corpus design in linguistics were laid by Douglas Biber. He notes:

10 Willard McCarty, 'Modeling: A Study in Words and Meanings', in *A Companion to Digital Humanities*, ed. by Susan Schreibman, Ray Siemens, and John Unsworth (Malden, MA: Blackwell, 2004), pp. 254–70 (p. 256).

11 Aleksej Vsevolodovič Gladkij and Igor Aleksandrovič Mel'čuk, *Elementy Matematičeskoj Lingvistiki* (Moscow: Nauka, 1969), p. 9.

12 Douglas Biber, 'Methodological Issues Regarding Corpus-Based Analyses of Linguistic Variation', *Literary and Linguistic Computing*, 5.4 (1990), 257–69 (p. 257).

13 Since most linguistic corpora are released in annotated form, the presence of annotations — which can range from general-purpose annotations such as part-of-speech tags to specific ones such as discourse roles or phrasemes — are sometimes taken as a feature distinguishing corpora from text collections. However, such annotations are a *further* modelling step, which logically *follows* the selection step. Our definition therefore does not require a corpus to be annotated.

> The use of computer-based corpora provides a solid empirical foundation for general purpose language tools and descriptions, and enables analyses of a scope not otherwise possible. However, a corpus must be 'representative' in order to be appropriately used as the basis for generalizations concerning a language as a whole […].[14]

Biber emphasizes that representativeness does not primarily depend on sample size; 'rather, a thorough definition of the target population and decisions concerning the method of sampling are prior considerations'.[15] This includes balancing the corpus for features such as genre, topic, and authors, diachronic and diatopic distribution, etc. Subcorpora, e.g. representing different genres or eras, are designed to be comparable by using the same number and size of samples. Even when one only wants to model formally published, written language, which — as opposed to everyday spoken language — is more or less documented in library catalogues, it is evident that a thorough definition of the target population is very hard.

In any case, it is here where the study of a linguistic phenomenon starts, with the sampling of utterances to construct a corpus that can serve as a model of language. This first step is obviously crucial: whatever is missing from the corpus, will also be missing from all subsequent analyses, and overrepresentation as well as underrepresentation will be hard to adjust afterward. Corpus design is thus of utmost importance for the validity of any results derived from the corpus.

The next step is then to look for evidence of the language phenomenon under research. This involves the construction of formal models, which can take many forms, but which can be thought of as 'queries' to the corpus. It is sometimes possible to formulate exact queries, in particular when the phenomenon is directly observable on the surface level, for example when one is interested in a specific word form. In other cases, additional annotation layers are needed, such as part-of-speech tags, lemmas, or morphological analyses. When dealing with historical texts (whether in a diachronic or synchronic fashion), it may be necessary to consider older spellings or syntactic constructions.

With respect to the use of corpora for historical research, Tony McEnery and Helen Baker note in their book that '[i]n spite of a large body of work in linguistics in general, and in corpus linguistics in particular, using corpora to explore the past is still in its infancy'.[16] In particular, while there do exist corpora of historical language, such as, for English, the Helsinki Corpus of English Texts and ARCHER (A Representative Corpus of Historical English Registers), 'the size and structure of those corpora reflect very much the types of research questions the corpora were designed to address',[17] namely the development of English grammar over a long period of time.[18] This should not come as a surprise: it merely illustrates the pragmatic aspect of corpora as models of language. Grammatical features and function words are more frequent than content words, so the size of the corpora[19] is sufficient to study, for example, the use of modals in English over time, but not to study seventeenth-century prostitution (the research question addressed by McEnery and Baker): the word *prostitute* occurs just twice in

14 Biber, p. 243.
15 Biber, p. 244.
16 Tony McEnery and Helen Baker, *Corpus Linguistics and 17th-Century Prostitution: Computational Linguistics and History* (London: Bloomsbury, 2017), p. 3.
17 McEnery and Baker, p. 4.
18 For an overview of historical linguistic corpora, see Piotrowski.
19 Helsinki: 1.5 million running word forms covering the period 730–1710; ARCHER: 3.2 million running word forms covering the period 1650–1959.

the seventeenth-century text samples in ARCHER and not at all in the corresponding slice of the Helsinki corpus.

McEnery and Baker therefore conclude: 'When we are investigating content words — where our focus is on meaning rather than grammar, as it is in this book — we need larger corpora than those available to date.'[20] In other words: for many linguistic purposes, one newspaper article is (within the construction criteria of the corpus) as good as another, and random sampling is thus a useful approach in linguistic corpus design. History, however, does not study language as such: in the study of most historical research questions, one cannot simply replace, say, the proceedings of the German Reichstag of 4 August 1914 with those of 21 March 1933, even though they could both be said to be examples of early twentieth-century German parliamentary proceedings. A *historical* notion of representativeness is obviously different from that of *linguistic* representativeness. This difference is of course due to the simple fact that the research objects of linguistics and history differ.

If we want to design a corpus that does not model language but something else, modelling (i.e. corpus construction) obviously needs to proceed differently. One approach is to sample on a different level of granularity: while linguistic corpora usually have target sample sizes defined in words (or rather *word forms*, to be precise) and sample on the level of sentences, one can also sample, for example, complete texts. In literary studies, a corpus aiming to model, say, 'the nineteenth-century novel' (for the purpose of studying genre, style, topics, plots, etc.) could thus be constructed by sampling, according to certain criteria, complete novels from the set of all novels published during the nineteenth century. This approach is eventually taken by McEnery and Baker by using Early English Books Online (EEBO) as the corpus for their research. This is also the approach taken by Laramée and by Daems and others in this volume.

The use of variable-size sampling units (novels may vary widely with respect to their length) may seem like a small adjustment in corpus construction, but it means that a number of fundamental assumptions no longer hold; for example, all frequency calculations need (at least) be normalized for length.

Serial Sources

Due to recent systematic digitization efforts, often initiated by libraries or (a successor of) the publisher or issuing institution, more and more historical periodicals and other serial publications are now digitally available in full, i.e. all of their issues, or at least all issues within a certain date range. Examples are parliamentary proceedings such as those of the German Reichstag (mentioned above), the journal *Die Grenzboten* (1841–1922) [Nölte and Blenkle, this volume], or Text+Berg, which contains all yearbooks of the Swiss Alpine Club (SAC) since 1864.[21]

Unlike the corpora discussed above, which are based on sampling from a larger set, these collections can be considered 'complete' in the sense that they contain, in fact, *all* items of a particular kind.

We have said before that corpora are models since they exhibit all three of Stachowiak's model properties, the mapping property, the reduction property, and the pragmatic property. Given that collections like Text+Berg are not reduced with respect to the issues they contain, and that the intended use and the intended users are usually

20 McEnery and Baker, p. 6.

21 Martin Volk and others, 'Challenges in Building a Multilingual Alpine Heritage Corpus', in *Proceedings of the Seventh International Conference on Language Resources and Evaluation*, ed. by Nicoletta Calzolari and others (Valletta: European Language Resources Association, 2010), pp. 1653–59.

left open at the time of digitization, are they corpora? In particular, since we have emphasized that it is the intentional *selection* of texts that defines a corpus, one could argue that the selection criterion in the case of these integral collections is trivial (in the mathematical sense of the word).

To answer this question, we need to consider several aspects. First of all, a digitization is certainly a model, like a photograph is a model; thus, Text+Berg is a model of the printed SAC yearbooks. Even though it contains (the text of) all issues, the model is reduced, for example, with respect to the properties of the paper, which are not mapped to the digital representation. Text+Berg is thus prima facie just a model of the printed SAC yearbooks; there is nothing in the selection criterion that would indicate otherwise. From a linguistic point of view, Text+Berg is certainly not a good model of general language, since it obviously covers only a small number of very specific types of utterances. Text+Berg *may* serve as a model for a smaller subset of language, say, the language of nineteenth-century mountaineering reports, but this is something that would need to be demonstrated — it does not automatically follow from the selection criterion.

If we think beyond language and linguistics, from a historical point of view periodicals can of course be considered models of discourses, ideas and ideologies, societal movements, trends, changes, and so on. Thus, even though a collection like Text+Berg may not be a good model of language, it may be a good model of some extralinguistic historical phenomenon; in this sense, such a collection may then also be considered a corpus. The question thus becomes whether a single periodical or a particular periodical is suitable and sufficient for modelling a particular phenomenon with respect to a particular research question. On the one hand, all of this is old hat to historians: after all, history could be said to be almost defined by using documents (sources) as models for studying the past, which itself is inaccessible. On the other hand, what is new is that such 'complete corpora' of periodicals are now available, easily accessible, and can be analyzed automatically — thus enabling historians to quickly perform analyses that used to be practically impossible before because they would have required too much time and manual labour.

These new possibilities are obviously welcome, but they also bring along new challenges. One of them is the allure of convenience, which may lead to opportunistically choosing a particular periodical as a 'model' for some phenomenon just because it is available as a corpus.

The same goes for the analyses: just because NLP methods and tools are available, this does not mean they are suitable for historical texts and research questions, as they were usually designed for completely different types of texts and research questions, and the underlying models are rarely made explicit. In addition, historical analyses of historical corpora are particularly challenging because (1) historical analyses usually take place on the level of meaning (i.e. semantics and pragmatics), and (2) historical language may differ significantly from modern language.

Semantic and pragmatic analyses are generally hard, but in particular in diachronic settings — as in the analysis of serial sources — where one has to be aware of meaning shifts and may thus need to consider seemingly unrelated surfaces. For example, in a collection of early modern German texts, a query for *Frau* ('woman') may be expanded to include historical spelling variants such as *frauw*, *fraw*, *frouw*, *frāw*, or *vrau*. Assuming that matching examples are in the corpus, this will indeed return results. However, these results may not be relevant, as historically *Frau* referred only to noblewomen; the historically equivalent word would be *Weib* (spelled *wib*, *wip*, etc.), which later acquired a pejorative meaning. Another, slightly different case is *Freund*, which today exclusively means 'friend' (in the sense of a companion), but which was a synonym for

Verwandter ('relative') until the early eighteenth century. Search results for a query of *Freund* may thus not actually be relevant, whereas a query for *Verwandter* will return only a small portion of the relevant results. Such problems obviously cannot be solved by the surface-level approaches that dominate modern NLP because they are sufficient for many commercial applications; for historical research, more semantic knowledge is required.

The increasing availability of historical sources — and serial sources in particular — in digital form holds a great potential for advancing historical research. Convenience is certainly one aspect: even if only scans are available, it is much quicker and easier to browse all issues of a journal on screen for potentially interesting articles than to visit a (potentially remote) archive or library and handle the physical items. The availability of digital full text promises a further acceleration: it is then possible to search and find instantaneously occurrences of names or keywords in huge collections. However, as quick and convenient as it is, it does not mark a methodological change. The actual potential of digitalization lies in the possibility to create computational models that make assumptions explicit, can be rigorously tested, revised, and shared in an agile manner. One important advantage is that other researchers can then also study and manipulate the models; for example, they could modify the queries, or test them on different corpora.

In any case, the corpus, its design (the underlying assumptions, selection criteria, metadata, types and sources of annotation layers, etc.), and the model of the phenomenon under study (instantiated by queries to the corpus) must be made explicit. Only then can the research be reproduced, and the advantages of formal models fully realized.

Conclusion

We have tried to make clear the importance of formal modelling for digital humanities; in fact, we argue that formal modelling is the core of digital humanities. A corpus can be considered as a model in the sense of Stachowiak: it is a reduced mapping of some original phenomenon for a particular purpose. The creation of a corpus is thus the construction of a model, and modellers consequently have to answer the questions: What is the original? In what respects is the model a reduction of it? And for whom and for what purpose am I creating the model? These are not new questions: every time historians select sources they construct models, even before any detailed analysis. With respect to large digital corpora, one could thus argue that the difference is merely a question of scale. But the larger and the more 'complete' a corpus is, the greater the danger to succumb to an 'implicit essentialism'[22] and to mistake the model for the original: it seems to contain 'everything', so it must be 'true' (something that can often be observed in the field of cultoromics,[23] when arguments are being made on the basis of the Google Books Ngram Corpus). The same obviously goes for any analysis of a corpus: if the corpus is 'true', so must be the analysis; if there is no evidence of something in the corpus, it did not exist. This allure is even greater when the analysis is done automatically and in particular using opaque quantitative methods: the computational analysis is neutral and objective, so there is no reason to question the results.

By their very nature, serial sources have a temporal dimension and thus lend themselves to the study of historical developments or, in other words, the construction of diachronic models. It is clear that the analysis of developments over time brings about a particular danger of teleological interpretations. However, computational models

22 Bernard Mothon, *Modélisation et vérité* (Paris: Archétype 82, 2010), p. 19.

23 Jean-Baptiste Michel and others, 'Quantitative Analysis of Culture Using Millions of Digitized Books', *Science*, 331.6014 (2011), 176–82.

can also help to counter this natural tendency, as they require assumptions to be made explicit. For example, suppose that browsing through the issues of a journal, one gets the impression that a particular topic becomes more important over time. How can importance of a topic be modeled? A straightforward model could be the frequency of one or more terms representing this topic. Even though this is a very simple (if not simplistic) model, a computational implementation already requires making numerous assumptions explicit, including the identification of sample period and size, a definition of the terms to be counted, and a specification of how occurrences are to be counted.

But this is not the end of the story: whether the frequency count supports the hypothesis or not, the construction of a computational model also — and perhaps most importantly — represents an opportunity to question the assumptions (e.g. the choice of terms) of the model, as well as the modelling *framework*. For example, as the first model was quantitative, what could a qualitative *countermodel* look like?[24] A particular strength of computational models in this context is that they can be easily manipulated and that they thus support exploratory approaches. The ability to easily change model parameters and to re-test the model helps to form a more comprehensive understanding of the phenomenon under research, but it also fosters 'scientific serendipity': even simply re-sorting a list or changing a visualization may yield unexpected findings.

If scholars want to use the potential of the computer as a modelling tool rather than just a mechanical tool to automate tedious work, they need to become aware both of their own models and of the models on which they rely, and make them explicit.

Michael Piotrowski is Professor of Digital Humanities at the University of Lausanne. He is also co-director of the Laboratory of Digital Humanities and Cultures of the University of Lausanne and academic co-director of the UNIL-EPFL Center for Digital Humanities. His research focuses on knowledge representation and formal modelling in the humanities and language technology for historical texts. He is the author of the first text book on Natural Language Processing for Historical Texts (2012), and also has a long-standing interest in document engineering and interactive editing and authoring aids (LingURed).

Bibliography

Biber, Douglas, 'Methodological Issues Regarding Corpus-Based Analyses of Linguistic Variation', *Literary and Linguistic Computing*, 5.4 (1990), 257–69

Bouleau, Nicolas, *La modélisation critique* (Versailles: Quæ, 2015)

Epstein, Joshua M., 'Why Model?' *Journal of Artificial Societies and Social Simulation*, 11.4 (2008), 12

Gershenfeld, Neil, 'Truth Is a Model', response to the thematic issue 'What Scientific Concept Would Improve Everybody's Toolkit?' *Edge* (2011), www.edge.org, no page

Gladkij, Aleksej Vsevolodovič, and Igor Aleksandrovič Mel'čuk, *Elementy Matematičeskoj Lingvistiki* (Moscow: Nauka, 1969)

Marczewski, Jean, *Introduction à l'histoire quantitative* (Geneva: Droz, 1965)

McCarty, Willard, 'Modeling: A Study in Words and Meanings', in *A Companion to Digital Humanities*, ed. by Susan Schreibman, Ray Siemens, and John Unsworth (Malden, MA: Blackwell, 2004), pp. 254–70

24 For a general discussion of countermodelling, see Nicolas Bouleau, *La modélisation critique* (Versailles: Quæ, 2015).

McEnery, Tony, and Helen Baker, *Corpus Linguistics and 17th-Century Prostitution: Computational Linguistics and History* (London: Bloomsbury, 2017)

Michel, Jean-Baptiste, Yuan K. Shen, Aviva P. Aiden, Adrian Veres, Matthew K. Gray, The Google Books Team, and others, 'Quantitative Analysis of Culture Using Millions of Digitized Books', *Science*, 331.6014 (2011), 176–82

Mothon, Bernard, *Modélisation et vérité* (Paris: Archétype 82, 2010)

Piotrowski, Michael, *Natural Language Processing for Historical Texts*, Synthesis Lectures on Human Language Technologies (Lexington, KY: Morgan & Claypool, 2012).

Stachowiak, Herbert, *Allgemeine Modelltheorie* [*General Model Theory*] (Vienna: Springer, 1973)

Volk, Martin, Noah Bubenhofer, Adrian Althaus, Maya Bangerter, Lenz Furrer, and Beni Ruef, 'Challenges in Building a Multilingual Alpine Heritage Corpus', in *Proceedings of the Seventh International Conference on Language Resources and Evaluation*, ed. by Nicoletta Calzolari and others (Valletta: European Language Resources Association, 2010), pp. 1653–59