Towards a quantitative research framework for historical disciplines

Barbara McGillivray
The Alan Turing Institute
University of Cambridge
bmcgillivray@turing.ac.uk

Jon Wilson
King's College London
jon.wilson@kcl.ac.uk

Tobias Blanke
King's College London
tobias.blanke@kcl.ac.uk

1 Background and motivation

The ever-expanding wealth of digital material that researchers have at their disposal today, coupled with growing computing power, makes the use of quantitative methods in historical disciplines increasingly more viable. However, applying existing techniques and tools to historical datasets is not a trivial enterprise (Piotrowski, 2012; McGillivray, 2014). Moreover, scholarly communities react differently to the idea that new research questions and insights can arise from quantitative explorations that could not be made using purely qualitative approaches. Some of them, such as linguistics (Jenset and McGillivray, 2017), have been acquainted with quantitative methods for a longer time. Others, such as history, have seen a growth in quantitative methods on the fringes of the discipline, but have not incorporated them into the mainstream of scholarly practice (Hitchcock, 2013).

Historical disciplines, i.e. those focusing on the study of the past, possess at least two characteristics, which set them apart and require careful consideration in this context: the need to work with closed archives which can only be expanded by working on past records (Mayrhofer, 1980), and the focus on phenomena that change in a complex fashion over time. First, that means historical research is grounded in empirical sources which are stable and fixed (one cannot change the archival record). But they are often hard to access and, recording the language and actions of only a small fraction of historical reality at any moment, have a complex relationship to the past being studied. Secondly, the categories through which the past is studied themselves change, making modelling, and the automation of analysis based on a limited number of features in the historical record a fraught enterprise.

Donald E. Knuth is maybe the most famous godfather of computer science. For him, "[s]cience is knowledge which we understand so well that we can teach it to a computer; and if we don't fully understand something, it is an art to deal with it. . . . [T]he process of going from an art to a science means that we learn how to automate something" (Knuth, 2007). Computing science is defined by the tension to automate processes using digital means and our inability to do so, because we fail to create fully explicit ways of understanding processes. In this sense, a computational approach to collecting and processing (historical) evidence would be a science if we could learn to automate it. Many features of the past can be understood through automation. Yet, the problematic nature of the relationship between sources and reality and the mutability of categories, means it will always rely on a significant degree of human intuition, and cannot be fully automated; computational history is an art in Knuth's terms.

The methodological reflections in this paper are part of an effort to think about how to define the possibilities and limits of quantification and automation in historical analysis. Our aim is to assist scholars to take full advantage of quantification through a rigorous account of the boundaries between science and art in Knuth's terms. Building on McGillivray et al. (2018), in this contribution we will begin with the framework proposed by Jenset and McGillivray (2017) for quantitative historical linguistics and illustrate it with two case studies.

2 A quantitative framework for historical linguistics

Jenset and McGillivray (2017)'s framework is the only general framework available for quantitative historical linguistics. A comparable framework,

but more limited in scope, can be found in Köhler (2012). Jenset and McGillivray (2017)'s framework starts from the assumption that linguistic historical reality is lost and the aim of quantitative research is to arrive at models of and claims on such reality which are quantitatively driven from evidence and lead to consensus among the scholarly community. The scope of application of this framework is delimited to the cases where quantifiable evidence (such as n-grams or numerical data) can be gathered from primary sources, typically in the form of *corpora*, i.e. collections of electronic text created with the purpose of linguistic analysis.

Jenset and McGillivray (2017) define *evidence* in quantitative historical linguistics as the set of "facts or properties that can be observed, independently accessed, or verified by other researchers" (Jenset and McGillivray, 2017, 39), and thus exclude intuition as inadmissible as evidence. Such facts can be pre-theoretical (as the fact that the English word *the* is among the most frequent ones) or based on some hypotheses or assumptions (as the fact that the class of article in English is among the most frequent ones, which is based on the assumption that the class of articles groups certain words together). *Quantitative evidence* is "based on numerical or probabilistic observation or inference" (Jenset and McGillivray, 2017, 39), and the quantification should be independently verifiable. On the other hand, *distributional evidence* has the form "x occurs in context y", where context can consist of words, classes, phonemes, etc. Annotated corpora, where linguistic (morphological, syntactic, semantic, etc.) information has been encoded in context, are considered as sources of distributional evidence to study phenomena in historical linguistics.

Following Carrier (2012), Jenset and McGillivray (2017, 40) define claims as anything that is not evidence, and statements are based on evidence or on other claims. The role of claims in the framework concerns their connection with truth, which can be stated in categorical terms (as in "the claim that x belongs to class y is true") or probabilistic terms (e.g. "x belongs to class y with probability p). Claims possess a strength proportional to that of the evidence supporting them. For example, all other things being equal, claims supported by large evidence are stronger than claims supported by little evidence.

Ultimately, research in historical linguistics aims at making (hopefully strong) claims logically following assumptions shared by the community, other claims, or evidence. A *hypothesis* originates from previous research, intuition, or logical arguments, and is "a claim that can be tested empirically, through statistical hypothesis testing on corpus data" (Jenset and McGillivray, 2017, 42). In this context, "model" means a formalized representation of a phenomenon, be it statistical or symbolic (Zuidema and de Boer, 2014). Models (including those deriving from hypotheses tested quantitatively against evidence) are research tools embedding claims or hypotheses, useful in order to produce novel claims and hypotheses in turn via "a continual process of coming to know by manipulating representations" (McCarty, 2004).

Based on these definitions, Jenset and McGillivray (2017) formalize the research process they envisage as part of their framework, see Figure 1. The process starts from the historical linguistic reality, which we assume to be lost for ever. Any research model can only aim at approaching this reality without reaching it completely, and quantitative historical linguistics ultimately will produce models of language that are quantitative driven from evidence. The rest of the diagram shows how this is achieved. The historical linguistic reality gave rise to a series of primary sources, including documents and other (mainly textual) sources, and these to secondary sources like grammars and dictionaries. Based on the knowledge of the language we gather from these sources we can draft annotation schemes which specify the rules for adding linguistic information to the corpora and thus obtain annotated corpora. Corpora are the source of quantitative distributional evidence which can be used to test statistical hypotheses, formulated based on our intuition of the language and on knowledge drawn from examples. Such hypotheses can also feed into the creation of linguistic models, which aim to represent the historical linguistic reality.

3 Model-building in history

In contrast with quantitative historical linguistics, the discipline of history possesses an extraordinary variety of idioms to describe itself, and has much less rigorous analytical vocabulary to describe its method. Yet there are important similar-

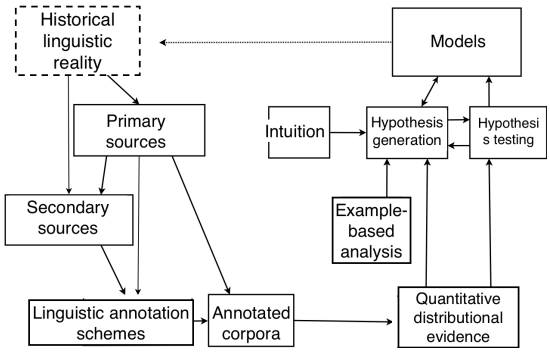

Figure 1: Research process from the quantitative historical linguistics framework described in Jenset and McGillivray (2017). Figure modified from Figure 2.1 in Jenset and McGillivray (2017, 45).

ities, which mean Jenset and McGillivray (2017)'s framework can be translated and modified for use for historical research more generally. First, historians assume that historical reality is lost, and can only be understood through traces left in a variety of archives (including human memory). Second, although historians rarely explicitly talk about constructing models, their practice largely consists of making claims about representations of the past which other disciplines would describe in precisely such terms. From the process they describe as the 'interpretation' or 'analysis' (Tosh, 2015) of the sources, historians create representations which reduce the vast complexity of historical reality to a few limited, stylised characteristics; Max Weber's Protestant Ethic, Lewis Namier's system of factional interest or C.A. Bayly's great uniformity. Third, these representations are used to make hypotheses and claims about change over time of different kinds. These might be about about the endurance or rupture of certain key feature in a particular sphere of activity, or about the forces responsible for causing a particular event or set of processes process, for example.

We have suggested that history is (if implicitly) essentially a model-building enterprise. That allows many of the hypotheses which historians develop to be theoretically amenable to quantification. The use of quantitative methods (in particular using the analysis of textual corpora) has increased recently (Guldi and Armitage, 2014). But, most historians are reluctant to quantify because they are skeptical about formalising their models, believing that to do so would imply their possessing a degree of categorical rigidity unwarranted by the complexity of the past. We suggest that more explicit reflection on method, and engagement with other fields (such as historical linguistics) which deal with fuzzy categories would help overcome these obstacles.

What's more, the use of digital data-sets and application of quantitative techniques to them allows historical claims based on the prevalence of certain features of the past to be empirically tested. Such claims are central to many forms of historical argumentation already; about the importance of particular concepts or practices at specific moments, for example. Of course such claims need to be precisely related to the structure of the (digitised) archive; as ever, limitations must be recognised. But given the amount of material which can be quickly processed, quantification allows claims previously asserted through little more the accumulation of anecdotes to be more rigorously validated.

4 Languages of power

The first case study where we apply Jenset and McGillivray (2017)'s framework considers a recent collaboration between Digital Humanities and History at Kings College London (Blanke and Wilson, 2017), to develop a "materialist sociology of political texts" following Moretti's ideas of distant reading (Moretti, 2013). The project worked on a corpus of post-1945 UK government White Papers to map connections and similarities in political language from 1945 to 2010. As the corpus is time-indexed, a quantitative analysis traced the changing shape of political language, by tracking clusters of terms relating to particular concepts and charting the changing meaning of words. Creating the distributional quantitative evidence involved text pre-processing to create a term-document matrix. Using natural language processing libraries, this was annotated with grammatical information, as well as with a number of dictionaries that reflected facets such as sentiment, ambiguity and so on. These allowed the project to use models for historical texts which not only read the texts themselves but also to developed ways of classifying them into time intervals. More advanced techniques were applied to trace changes of meaning in key political concepts across time intervals, using topic models and word embeddings, allowing historiographical and linguistic hypotheses to be tested.

In Jenset and McGillivray (2017)'s terms, these various techniques produced a variety of different quantitative distributional evidence, which allowed a series of hypotheses to be developed and tested. Intuition, often developed from historical research using non-quantitative techniques, had an important role in framing hypotheses. But quantitative evidence was able to impart greater clarity and specificity to intuitional hypotheses, often closing down multiple possibilities. For example, using our dictionaries demonstrated a major break in the language of White Papers in the mid-1960s, around the election of Harold Wilson's Labour government. While this intuitionally made sense, so would a break in the early 1980s, which we did not find, instead seeing a rupture in the early 1990s.

Combining our chronological analysis with topic modelling and word embeddings allowed us to build a series of models of the predominant concerns and the structure of political language in each epoch. In line with In Jenset and McGillivray (2017)'s framework, these models were built from iteratively generating and testing hypotheses. For example, we tested the frequency of different term clusters generated through topic modelling, and the terms whose embedding changed most dramatically between each epoch.

Our process of hypothesis generation and testing always had in mind the commonplace assumptions made by historians using non-quantitative techniques in the field. In many respectives, quantitative distributional evidence produced hypotheses at variance with those scholarly norms. For example, we found White Papers in the period from 1945 to 1964 to be dominated by post-war foreign policy concerns, not the construction of the welfare state; economic language was being dominant in the period from 1965-1990 not afterwards; and 'the state' as a political agent is more important in the later period than before.

Yet, as challenging as they may be to much of the historiography of post-war Britain, the form of these hypotheses is very similar to the form of the claims made in standard historical argumentation; there is no dramatic epistemological leap in the type of knowledge being produced. Although our models were developed using automated techniques, they can be verified qualitatively in the same way as non-quantifiable claims, through quotation, and the interpretation of words and phrases in specific contexts.

One important finding is the need to recognise the broad range of different ways in which quantitative analysis can be expressed. It is important, for example, to indicate the absolute frequency of terms in any series as well as their relation to other terms. There is significant work to be done developing ways to visually represent the quantitative features of any corpus of texts.

5 Predicting the Past

Digital humanities generally use computational modelling for exploratory data analysis. Digital humanities makes use of the advancements in the abilities to visualise and interactively explore in a relatively free fashion. Recently, we have witnessed the emergence of new combinations of exploratory data analysis with statistical evidence for discovered patterns. In the digital humanities, this is popular, too, if Klingenstein et al. (2014), for instance, integrats a historical regression analysis into their data visualisations. Our first example above is an instance of exploratory data analysis, using topic modelling and other tools to provide statistical evidence for underlying trends in the documents, as earlier demonstrated. Models, however, often have another purpose beyond the exploration of data. They are part of predictive analytics. Abbott (2014) is one of the most famous practitioners in the field. For him, predictive analytics work on "discovering interesting and meaningful patterns in data. It draws from several related disciplines, some of which have been used to discover patterns in data for more than 100 years, including pattern recognition, statistics, machine learning, artificial intelligence, and data mining." (Abbott, 2014).

It is a common misunderstanding to reduce predictive analytics to attempts to predicting the future. It is rather about developing meaningful relationships in any data. Predictive analytics compared to traditional analytics is driven by the data under observation rather than primarily by human assumptions on the data. Its discipline strives to automate the modelling and finding patterns as far as this is possible. In this sense, it moves away from both exploratory and confirmatory data analysis, as it fully considers how computers would process evidence.

O'Neil and Schutt (2013) introduce the idea of predicting the past, which is used to model

the effects of electronic health records (EHR) and to set up new monitoring programs for drugs. For O'Neil and Schutt (2013), these integrated datasets were the foundations of novel research attempts to predict the past. They cite the 'Observational Medical Outcomes Partnership (OMOP)' in the US that investigates how good we are at predicting what we already know about drug performance in health using past datasets. Once OMOP had integrated data from heterogeneous sources, it began to look into predicting the past of old drug cases and how effective their treatments were. "Employing a variety of approaches from the fields of epidemiology, statistics, computer science, and elsewhere, OMOP seeks to answer a critical challenge: what can medical researchers learn from assessing these new health databases, could a single approach be applied to multiple diseases, and could their findings be proven?" (O'Neil and Schutt, 2013). Predicting the past thus tries to understand how "well the current methods do on predicting things we actually already know" (O'Neil and Schutt, 2013).

Such a novel approach relating to past data sets should be of interest to the digital history. Digital history could use the approach to control decisions on how we organise and divide historical records. An existing example that implies predicting past events by joining historical data sets, is the identification of historical spatio-temporal patterns of IED usage by the Provisional Irish Republican Army during The Troubles, used to attribute historical behaviour of terrorism (Tench et al., 2016).

In Blanke (2018), we demonstrate how predicting the past can complement and enhance existing work in the digital humanities that is mainly concentrated on exploring gender issues as they appear in past datasets. Blevins and Mullen (2015) provide an expert introduction into why digital humanities should be interested in predicting genders. Gender values are often missing from datasets and need to be imputed. Predictive analytics can be seen as a corrective to existing data practices and we can predict the genders in a dataset. In Blanke (2018), we compare a traditional dictionary-based approach with two machine learning strategies. First a classification algorithm is discussed and then three different rule-based learners are introduced. We can demonstrate how these rule-based learners are an effective alternative to the traditional dictionary-based method and partly outperform it.

Blanke (2018) develops the predicting the past methodology further and present differences from other predictive analytics approaches. We follow all the steps of traditional predictive analytics to prepare a stable and reliable model, where we pay particular attention to avoid overfitting the data, one of the main risks in predictive models. An 'overfitting' model is one that models existing training data too closely, which negatively impacts its ability to generalize to new cases. We perform extensive cross-validations to avoid over-fitting.

Predicting the past, however, differs significantly from other approaches, as the model is not prepared for future addition of data but to analyse existing data. The aim is to understand which (minimal) set of features makes it likely that observation x includes feature y. In Blanke (2018), we aimed to understand which combination of features make it likely that a historical person is of gender female, male or unknown. The next step in our methodology is therefore to apply the best performing models to the whole data set again to analyse what gender determinations exist in the data. Is it, e.g., more likely that vagrants were female in London?

The common approaches to gender prediction in the digital humanities uses predefined dictionaries of first names and then matches the gender of individuals against this dictionary. This has firstly the problem that these dictionaries are heavily dependent on culture and language they relate to. But this is not the only issue, as dictionary-based approaches secondly also assume that errors are randomly distributed. Gender trouble is simply a problem of not recording the right gender in the data. Our predictive analytics approach in Blanke (2018) on the other hand does not make this assumption in advance and judges gender based on the existing data. This has led in turn to interesting insights on why certain genders remain unknown to the models.

In summary, predicting the past is based firstly on going through all traditional predictive analytics steps to form a stable model that reflects the underlying historical evidence close enough but also does not overfit. Secondly, we use this stable model to algorithmically analyse historical evidence to gain insights on how a computer would see the relations of evidence.

6 Conclusion and future work

This comparison leads us to the conclusion that, despite the broad applicability of Jenset and McGillivray (2017)'s framework in both cases, some important differences emerge between historical linguistics and history. We discuss two. First of all, the scope of primary source and its quantitative representation is broader in history, including not only distributional but also categorical, ordinal, and numerical evidence. History requires careful discernment of which is most appropriate, and how they should be combined.

Secondly, the scope for a purely quantitative approach is less broad: quantitative evidence and models can often only contribute to inform hypotheses and claims which rely on qualitative evidence and methods. Often it seems that quantitative methods are only accepted by historical scholars if the claims developed by automated techniques can also be verified qualitatively, through anecdote, quotation and so on. In many fields quantification can be accepted because it creates results which look similar to those produced by qualitative research. But this approach limits the development of methods that use quantification to do more than simply re-frame qualitative observations, and instead make statistical arguments about aggregate behaviour in its own right. In the future, we plan to develop these insights further, in order to build a more comprehensive research framework which integrates qualitative and quantitative approaches.

Acknowledgments

This work was supported by The Alan Turing Institute under the EPSRC grant EP/N510129/1. BM is supported by the Turing award TU/A/000010 (RG88751).

References

Dean Abbott. 2014. *Applied predictive analytics: Principles and techniques for the professional data analyst*. John Wiley & Sons.

Tobias Blanke. 2018. Predicting the past. *DHQ: Digital Humanities Quarterly* 12(2).

Tobias Blanke and Jon Wilson. 2017. Identifying epochs in text archives. In *2017 IEEE International Conference on Big Data (Big Data)*. pages 2219–2224.

Cameron Blevins and Lincoln Mullen. 2015. Jane, john... leslie? a historical method for algorithmic gender prediction. *DHQ: Digital Humanities Quarterly* 9(3).

Richard Carrier. 2012. *Proving history: Bayes's theorem and the quest for the historical Jesus*. Prometheus Books, Amherst, N.Y.

Jo Guldi and David Armitage. 2014. *The History Manifesto*. Cambridge University Press, Cambridge.

Tim Hitchcock. 2013. Confronting the digital: Or how academic history writing lost the plot. *Cultural and Social History* 10(1):9–23.

Gard B. Jenset and Barbara McGillivray. 2017. *Quantitative Historical Linguistics. A Corpus Framework*. Oxford University Press, Oxford.

Sara Klingenstein, Tim Hitchcock, and Simon DeDeo. 2014. The civilizing process in londons old bailey. *Proceedings of the National Academy of Sciences* page 201405984.

Donald E Knuth. 2007. Computer programming as an art. In *ACM Turing award lectures*. ACM, page 1974.

Reinhard Köhler. 2012. *Quantitative Syntax Analysis*. de Gruyter Mouton.

Manfred Mayrhofer. 1980. *Zur Gestaltung des etymologischen Wörterbuchs einer "Großcorpus-Sprache"*. Akademie der Wissenschaften. Phil-Hist. Klasse., Wien: Österr.

Willard McCarty. 2004. Modeling: A study in words and meanings. In Susan Schreibman, Ray Siemens, and John Unsworth, editors, *A Companion to Digital Humanities*, Blackwell Publishing Ltd., Malden, MA, USA, pages 254–270.

Barbara McGillivray. 2014. *Methods in Latin Computational Linguistics*. Brill, Leiden.

Barbara McGillivray, Giovanni Colavizza, and Tobias Blanke. 2018. Towards a quantitative research framework for historical disciplines. In *COMHUM 2018: Book of Abstracts for the Workshop on Computational Methods in the Humanities 2018*. Lausanne, Switzerland.

Franco Moretti. 2013. *Distant Reading*. Verso, London.

Cathy O'Neil and Rachel Schutt. 2013. *Doing data science: Straight talk from the frontline*. " O'Reilly Media, Inc.".

Michael Piotrowski. 2012. *Natural language processing for historical texts*. Morgan & Claypool, San Rafael, CA.

Stephen Tench, Hannah Fry, and Paul Gill. 2016. Spatio-temporal patterns of ied usage by the provisional irish republican army. *European Journal of Applied Mathematics* 27(3):377–402.

John Tosh. 2015. *The Pursuit of History. Aims, Methods and New Directions in the Study of History*. Routledge, London, sixth edition.

Willem Zuidema and Bart de Boer. 2014. Modeling in the language sciences. In Robert J. Podesva and Devyani Sharma, editors, *Research Methods in Linguistics*, Cambridge University Press, Cambridge, pages 428–445.