The Death of the Historical Corpus

Laurent Gauthier

The Death of the Historical Corpus*

Laurent Gauthier†

This version: August 28, 2021.

Abstract

Although historians intuitively know what a corpus is, and exploiting a corpus is recognized as central in historiography, there has been little methodological focus on coming to a unified approach to the design and use of corpora. The massive expanse of digital information and processing capabilities over the past few years has also led to a diversity of approaches. After reviewing the history of the use of corpora in historiography, we examine how historians have taken possession of digital practice, and how it has interacted with the notions and uses of corpora: there are many diverse and somewhat incompatible perspectives. Next, we show how the corpus, as an input into historiography, can not exist anymore as an object and must be seen as a process or a pipeline. Then, its multiple and sometimes opposite perceptions can be unified, at the same time making history more scientific in the sense of Lucien Febvre's definition.

Keywords: Digital humanities, historiography, historical corpus

*__Draft working paper__

†Email: laurent.o.gauthier@gmail.com. Laboratoire d'Économie Dionysien, Université Paris 8 Saint-Denis Vincennes, EA 3391 - Bâtiment D, 2, rue de la Liberté, 93526 Saint-Denis, France.

All historians, from classicists to cliometricians, know what a corpus is in historiography, and they continuously manufacture new ones. Still, the features of an underlying corpus are not always explicitly mentioned as such in every research publication, and the term remains quite polysemic; referring to a corpus does not conjure up a single unequivocal definition. For example, it may at times designate a large set of texts from which one will isolate excerpts of interest, or it may designate the excerpts, or it may designate some serialized information derived from such excerpts. In spite of the fuzziness surrounding the concept, establishing a corpus in order to be able to rely on a specific set of documentary evidence has been deemed absolutely central in the historiographical process. This tension between necessity and a lack of precise determination is aggravated by the availability of an ever increasing volume of digitized information, which by its nature demands sorting and categorization.

Hence, it is no wonder that methodological papers on the use of digital tools have not converged in any way, if not to say that historians rarely used the digital environment up to its full potential. Applying exponential power to an imperfectly defined object could not result in broad agreement on a methodology. Here, we will argue that, once we account for digitization, the corpus in historiography does not exist anymore as an object. The corpus is not relevant as an ontology, as a body, but instead as *anima*, as an abstract process in the sense of information theory and computer science.

In order to reach this conclusion, we will first examine the history of the use of a corpus in historiography, up to some recent questioning of historians' methods. Having established its centrality, we will look into the reception of digital methods and their interaction with the notions and uses of corpora. Finally, we will see that these multiple and sometimes opposed views on the embedding of corpora in the digital context can be aligned if we consider the corpus a process, and not an object. We will see what changing the perspective translates into in terms of practice, and propose a few recommendations.

A Brief History of the Corpus

Current perspectives on historiography see a continuum between the archive, the document, the source and the corpus, and all essentially recoup with the notion of "documents produced by the actors of the history under study[1]". Since all historians have a clear understanding of what a corpus is, one could expect there would be a strong trend towards defining and clarifying the

[1] "les documents produits par les acteurs de l'histoire étudiée", see Offenstadt (2011), p. 68.

notion. That has not been the case: for example, it has been noted that the use of the term *corpus* among French medievalists is widely spread, but that there has not been any reflection whatsoever about its meaning nor any effort at conceptualization[2]. Before delving into the history of corpora in historiography, we therefore propose a simple and temporary definition in order to avoid any ambiguity: a *corpus* is a set of documents[3].

We can shed some light on the historians' actual use of the concept of corpus by serializing textual data on research articles and books in various academic fields. We use the "Data for Research" tools that have been developed for this purpose by JSTOR[4]. Figure 1 displays the percentage of research publications that contained the word *corpus*, through the 20th century and up to 2021. One important realization is that in some fields, such as mathematics and physics or social and political sciences, the use of the term has remained generally low (even declining in mathematics and physics), but it rose in history and in linguistics. Writing of corpora is therefore not a general trend across all scientific fields, and is quite specific to linguistics and history. The development of corpus linguistics after the 1950s can indeed be read directly on this chart. Interestingly, the term was slightly more in use in history in the 1920s and 1930s than in linguistics. In the 21st century, its use has accelerated in history, while it plateaued in linguistics. The facts that the corpus has a particular use in history, as opposed to various other fields, and that this use is becoming almost as common as in linguistics, are significant.

In order to contextualize the ever-increasing presence of the notion of corpus in historiography, we will first briefly trace the historical roots of its use, and then examine how historians and others have recently questioned the use of this concept.

The Development of Corpora in History

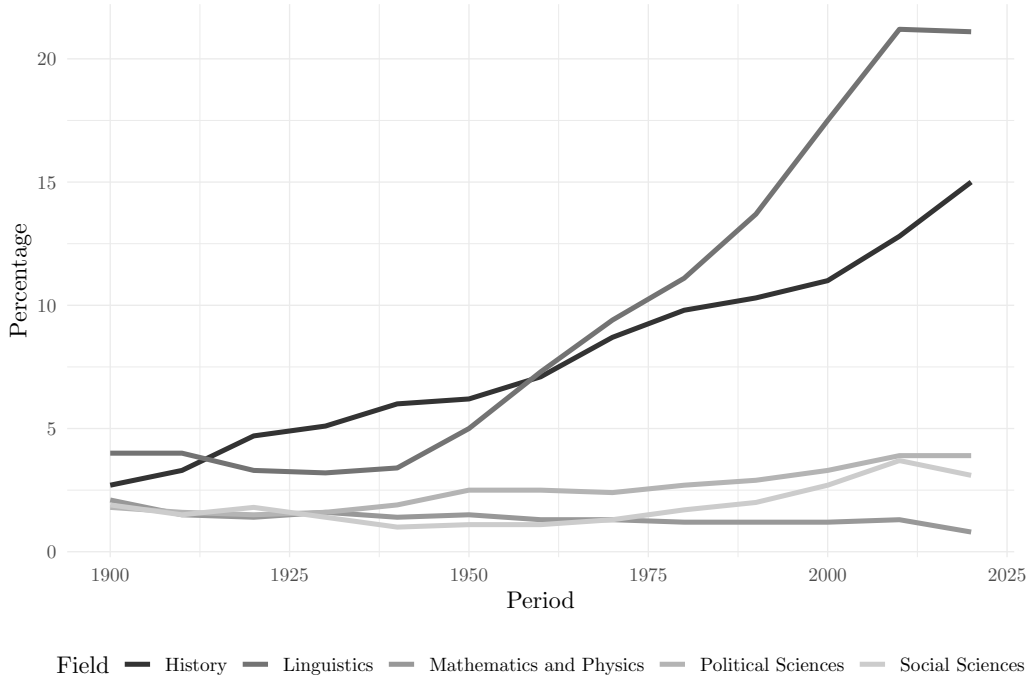
The notion of document and the writing of history have been associated for a long time. Constituting a corpus to carry out historical analysis is part and parcel of historical work and has even been considered to sufficiently define the very fact of writing history. In 1934, in a violent critique of a monograph on royal accounting, which triggered an intense historiographical debate, Lucien Febvre insisted that analysis was more important than erudition. He still recognized that putting together a corpus defined the writing of history, "the most exacting attention to

[2]See Magnani (2017a).

[3]Interestingly, recent work interrogating the notion of corpus in history considered its etymology but did not unequivocally define the term: see Treffort (2014) and Magnani (2017a).

[4]See for example Burns et al. (2009), as well as JSTOR's website for up-to-date information.

Figure 1: *Percentage of JSTOR Articles, Books or Book Chapters in English Containing the Term* corpus*, by Academic Field*



procuring usable documents, unfold them, file them, and sort them in a rational order[5]". As much as a corpus defines the writing of history, it sometimes even exists in an organic relationship with the historian: Arlette Farge described how the materiality of the archive played an important role in her writing process[6].

How has history writing come to rely on constituting corpora as the centerpiece of its methodology, "la centralité de l'archive" as Jean Boutier coined it[7]? As Jose Carlos Bermejo Barrera pointed out in his historiographical study, there could be no notion of a corpus or document in ancient historiography, because the very witnessing of events served as the foundation for history writing[8]. Relying on one's direct witnessing of events naturally implies that history writing was mainly concerned with contemporary history, and texts were held as inferior witnesses[9]. Boutier also distinguishes the historians, who wrote about contemporary events, from antiquists, such as Varro, who relied on a close study of text and artifacts to examine prior customs and ways of life[10], and both endeavors remained distinct until the end of Antiquity.

[5] "L'attention la plus exacte à se procurer les documents utilisables, à les dépouiller, à les mettre en fiches, à classer ces fiches dans un ordre rationnel." See Febvre (1934), p. 149.

[6] She manually recopied many 18th century judicial archives *in extenso*, as she explained in Camar et al. (2017).

[7] See Boutier (2014), p. 10-11.

[8] See Bermejo Barrera (2001), p. 193.

[9] See Boutier (2014), p. 12 and p. 14.

[10] Boutier (2014), p. 15.

It is only when the history of the Church became a domain of inquiry that historians began to consider the existence and necessity of a text corpus. This body of text had a materiality, just like sacred relics did, and the direct testimony from the contemporaries of Christ could not carry as much importance as personal testimony could have in the ancient Greek tradition[11]. Close reading of the ancient texts developed into hermeneutics, at the very foundation of the 18th and 19th century crystallization of a historical method. The first university seminar focusing on the examination of primary sources took place at the university of Göttingen in 1766[12]. According to Gunther Pflug, this prevalence of the document at first prevented any form of deductive operation, and in the 18th century thinker Pierre Bayle's perspective: "The scholar's goal consisted of surveying the factual data, penetrating the historical givens, without attempting to impose any order unless it were for mere purposes of clarity[13]."

Voltaire and later Turgot pulled history away from straight facts towards scientifically inspired analysis, relying on reason and common sense, thereby making the notion of pure document-based facts less central. Still critical and essential, the corpus now functioned with the application of reason, and inserted itself in the context of the question asked by the historian[14]. As the writing of history became professionalized in the 19th century, the methodology of source critique converged towards current practice[15]. At that juncture, historical knowledge acquired "a new configuration thanks to the introduction of two notions: that of document [...]; and that of the scientific method[16]." This perspective effectively established the document and the aspiration to a scientific approach as two facets of the same coin. Indeed, in a Foucaldian approach, analysis that is specifically historical as well as a more globalized form of analysis common in the social sciences at large both stem from the same source: "the questioning of the *document*[17]". Bermejo Barrera stresses the primordial place that the corpus holds in current historiography: "History builds its object starting from the constitution of its documentary corpora; it then develops different methods of reading and interpreting the texts, methods that are sometimes contradictory and that are not reducible to a common factor[18]."

This view largely recoups with Foucault's perception of historiography's position with respect

[11]Bermejo Barrera (2001), p. 194.

[12]Boutier (2014), p. 19.

[13]Pflug (1971).

[14]See Pflug (1971), p. 9-12 and p. 20-21.

[15]Offenstadt (2011), p. 70.

[16]Bermejo Barrera (2001), p. 198.

[17]"la mise en question du *document*", see "Foucault (1969), p. 13.

[18]Bermejo Barrera (2001), p. 204

to the document: it seeks not to interpret it, but to work it from the inside and elaborate it[19]; hence the document should not to be seen as inert material. Foucault defines the writing of history as the manner in which a mass of documents is organized[20].

It hence follows that the creation of a coherent corpus is one of the salient issues in historical methodology: determination of the documents to include or not, determination of the level of detail of the analysis that is required, determination of an analytic method for the quantification of various aspects from the documents[21]. Recognizing the influence that the presence or absence of a particular document may have on the writing of history, Boutier raised an important issue: one may ask which document should be used for which inquiry, but one should also ask to what extent the historian's questions will drive the gathering of a corpus[22]. The use of a corpus in historiography, for Foucault, strongly recoups with a necessary serial perspective: it is constituted following a particular and systematic methodology, and lends itself to quantitative analyzes[23]. Collecting similar or comparable elements, by construction, creates information that can be processed in a serialized form. Magnani further characterizes the corpus in a historical context as a "controlled observation", conjuring up the notion of an experiment that was repeated multiple times, from which one obtains a series of results[24].

Questioning the Notion of Corpus in History

The line for the field of history in Figure 1 showed a progressive increase, and the particular ideas and perspectives over the past hundred years or so we pointed out above hence percolated through the domain in a continuous manner. It is presumably in reaction to this accelerating trend that some historians have recently sought to put their practices into question as they pertain to the notion of corpus. While there is no question that the document is central in historiography, the issue is raised to precisely establish what a corpus is, and how one should make one. Implicitly following Foucault, Cécile Treffort stresses that a corpus has meaning, an *anima* inspired by the historian, and in that sense would differ from a simple set of documents,

[19] "[L'histoire] a changé sa position à l'égard du document : elle se donne pour tâche première, non point de l'interpréter, non point de déterminer s'il dit vrai et quelle est sa valeur expressive, mais de le travailler de l'intérieur et de l'élaborer [...]", see Foucault (1969), p. 14.

[20] "l'histoire, c'est une certaine manière pour une société de donner statut et élaboration à une masse documentaire dont elle ne se sépare pas.", Foucault (1969), p. 14.

[21] Foucault (1969), p. 19.

[22] See Boutier (2014), p. 10.

[23] Foucault (1969), p. 19.

[24] "le corpus est la configuration matérielle propre à une observation contrôlée, une réification formalisable de l'intertextualité qui serait insaisissable autrement." see Magnani (2017a), p. 5.

it "emanates from, translates and illustrates the researcher's thought[25]". At the same time, the corpus is tasked with aiming at comprehensiveness or at least representativity for the question under study. It therefore appears that one could distinguish the most comprehensive corpus containing every available document, which we could name the **universal corpus**, from the one specifically formed as a subset of the larger corpus in order to address a specific question, which could be termed the **oriented corpus**. This distinction recoups to a large extent with that between a corpus with a collective aim (made available online, for example) and a corpus with a personal aim (for the researcher), drawn by Treffort. Magnani pointed out that the use of the term by French medievalists was very often qualified with a possessive[26], which stresses the personal and specific manner in which the corpus is constituted, from the standpoint of the historian.

In an effort to combine perspectives on the corpus in linguistics, literature and human sciences, Damon Mayaffre proposed a conception of the corpus that reflected the ongoing interaction between the final interpretative act of a corpus and the original act of its constitution[27], which in effect internalizes the tension that Boutier pointed out. First, one needs to realize that creating a corpus necessarily implies a serial view, as Foucault had already mentioned. Second, the corpus, which in this context must be an oriented corpus as a research object, is an arbitrary construction whose worth only comes from the questions it raises and the answers it provides[28]. Mayaffre insists on these two important and separate steps in research, the constitution of the corpus, and the design of a treatment process for that corpus. Doing so, he raises the issue of the delineation of the corpus: in order to analyze a particular set of texts, one may have to resort to references outside this set, and this fundamentally constitutes an extension of the corpus, which may not be as thoughtful as the original one and stay at the level of the intuition[29]. This clearly poses the question of the proper definition of an oriented corpus out of a universal corpus, and the non-absolute manner in which it may be carried out. Mayaffre proposes the notion of a **reflective corpus**[30], which we can express as a corpus such that its constituents make up a semantic network, implicitly with an oriented corpus as its kernel. In fact, this perspective on building a corpus puts intertextuality at its core. Hence, from a universal corpus, one may

[25] "le corpus documentaire de tout chercheur est tout à la fois l'émanation, la traduction et l'illustration de sa pensée." See Treffort (2014).

[26] For example "mon corpus", "notre corpus", see Magnani (2017a).

[27] See Mayaffre (2002).

[28] Mayaffre (2002), p. 3.

[29] Mayaffre (2002), p. 5.

[30] "un corpus aujourd'hui, par sa taille et la compilation souvent exhaustive des textes afférents à un domaine donné, gagne à être *réflexif*", Mayaffre (2002), p. 5.

constitute a first oriented corpus seeking to address a particular research question, out of which one would grow a reflective corpus. However, by asking historians to potentially consider every possible extension of their corpus, the reflective corpus's superset, that is the universal corpus from which the oriented corpus is constituted, could be held to contain every text in the language or on the subject of interest[31]. Therefore, the corpus in history must be seen as something flexible and evolving, whose limits are dependent on the historical question and on the researcher's own discretionary decisions.

In light of this historical perspective, we conclude by stressing the crucial importance of the dialog between the constitution of a corpus and its effective exploitation. As the power of the digital processing of texts advanced, how have these principles actually been put to practice?

The Digital and the Historical Corpus

It may seem as though the introduction of digital tools has had little impact on historians' practices. For Philippe Rygiel, historians are "hypertextual polygraphs, who dissimulate most of the incriptions they produce[32]", stressing the fact that the majority of the historian's work is not visible from the results or the analyzes they publish. In Rygiel's view, the historian's annotations, essentially in textual form, constitute the core of their work, their production. In this perspective, the historical inquiry becomes the delineation of a corpus, augmented with these annotations[33]. The production of annotations is organically linked with the material corpus off of which it is based, and we can see how this would logically end up in a possessive attribution, as Magnani pointed out. How does the digital impact this framework? Rygiel observes that this historian production, the annotations, are in practice made in a digital framework and hence should ideally be made accessible to all, but they are generally not. One reason for this, he argues, is that if this production was systematically made public by providing all with the same raw material, it would end up rising the bar of expectations among all historians[34].

Focusing on many of the tools that are nowadays available to historians for textual analysis (such as Google keyword searches or newspapers electronic archives), Tim Hitchcock pointed out that "academic historians did not ask for these resources, and nor for the most part have they been

[31]This principle could logically be extended so that the whole bibliography one uses is part of the corpus, see Mayaffre (2002), p. 7.

[32]"L'historien contemporain apparaît alors d'abord comme un polygraphe hypertextuel dissimulant aux regards l'essentiel des inscriptions qu'il produit", see Rygiel (2011), p. 32.

[33]Rygiel (2011), p. 34.

[34]Rygiel (2011), p. 38.

directly responsible for their creation[35]". These perspectives seem to establish that the evolution towards digital frameworks, tools and analyzes has taken place in spite of, rather than thanks to, historians. Hitchcock's observations recoup those of Rygiel: he notes that the manner in which historians carry out their work has not changed substantially from the 1980s.

Some have argued that on the contrary, the very act of publishing could massively benefit from the new tools afforded by a purely digital framework: Nawrotzki and Dougherty illustrated this idea quite efficiently by using a collaborative and web-based publication process for their book on the subject[36]. They stress that, in their experience, the opening and sharing of resources did not increase competition between scholars, but rather led to more collaboration.

As we can see, in theorizing the use of digital tools by historians, the focus has tended to remain more on the process of writing history than on the objects off of which it is written: the historian rather than the corpus. Nevertheless, we will observe that, in fact, the use of digital tools for corpus manipulation has grown over time; we will first try to measure this evolution, and just as importantly, to compare it with other related concepts. Then, we will examine certain trends in the manner in which digital corpora are created or used, focusing on recent methodological research as well as on some examples of how digital methods were implemented by historians.

A Perspective on the Evolution of Digital Uses

As Gibbs and Owens stated a few years ago, "historical scholarship increasingly depends on our interaction with data[37]" and indeed such a statement would probably match most historians' intuition. Beyond simple intuition, we can gauge the growing importance of the use of, and references to, digital methods with Figure 2. The chart shows, for research publications in history and in English, the share of documents that contain some terms of interest[38]. The terms *data*, *analysis* and *corpus* are useful as controls, allowing us to better contextualize the evolution of references to *digital*. There is, in the curves for *data* and *analysis*, a strikingly similar hump from the 1950s to the 1980s: a likely reflection of the contemporary success of serial approaches, the French *Annales* school's influence, and a more spread-out focus on economic history at the time. Both series also exhibit an acceleration over the past 10 years or so, which could be related to the exponential rise in the use of the term *digital* since the turn of the century. The chart also

[35]See Hitchcock (2013), p. 10.

[36]See Nawrotzki and Dougherty (2013).

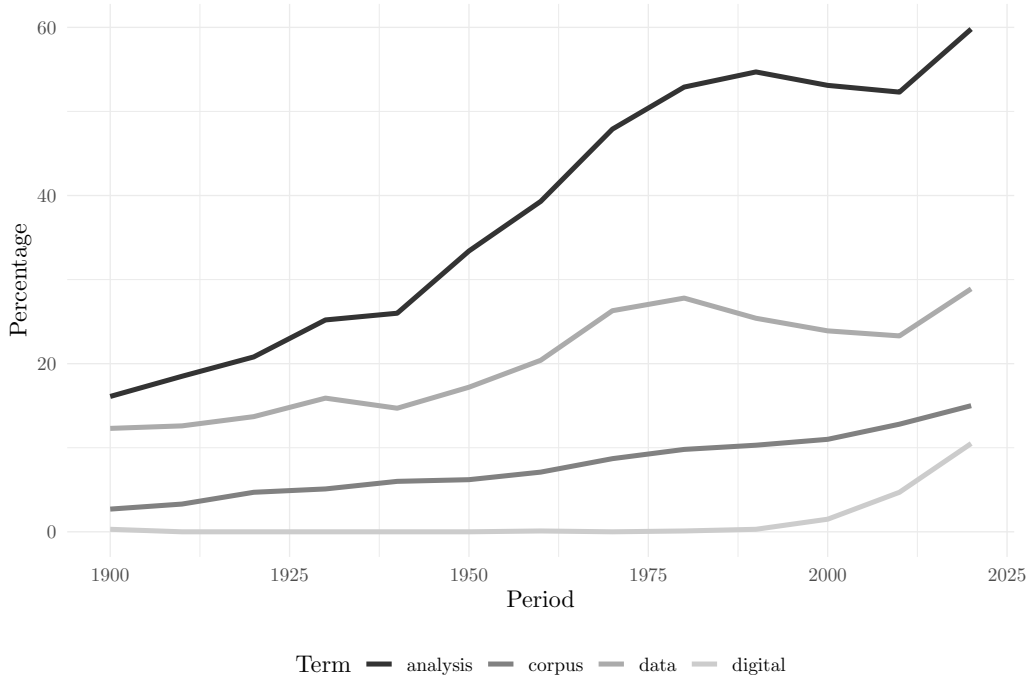
[37]See Gibbs and Owens (2013), p. 159.

[38]This analysis makes use again of the "Data for Research" feature in JSTOR. Note that the values for the line for *corpus* are the same as those in Figure 1, only the scale differs.

allows us to moderate the perception of the increase in the use of *corpus* we had noted earlier: in fact, it is dwarfed by the increase in historians' use of *digital*, *data* and *analysis*.

The continuous and long-term trend whereby historians increasingly resort to the notion of corpus in framing their work has hence recently (in the past ten or twenty years) been dominated by digital and data-driven approaches. That is not to say that quantitative analyzes are more prevalent, only that the writing of history has more and more relied on electronic sources, from whose perspective text naturally equates with data.

Figure 2: *Percentage of JSTOR History Articles, Books or Book Chapters in English Containing Specific Terms*



As the term *digital* gained importance in historiography, it has been associated with that of *corpus* in many recent considerations on historical analysis. In her review of the use of the notion of a corpus by medievalists, Magnani took digital processing as a given, and in his definition of the notion of reflective corpus, Mayaffre proposed that it be in practice structured as hypertext, whereby each text in what we called the oriented corpus would be linked to the its parent texts, in particular using standard XML encoding to account for these connections, implicitly stressing the necessity of an electronic representation of the corpus[39].

[39]See Mayaffre (2002), p. 8.

Viewing Digital Corpora as Databases

In many cases, historians approach the creation of a corpus specifically under the guise of an ontology (in the sense of information theory), that is a formal representation of knowledge, typically in the form of a database, which forms a structured view of the data and of some of its relations. In order to build a corpus for her dissertation, Ansley Erickson created a database as a way of keeping track of her notes relating to particular sources or material[40]. This led her to reflecting upon the role of categorization in storing data for a historiographical use. Attaching attributes to the data that was thus created (as opposed to, for example, a simple alphabetical organization) allowed for more flexible thought processes pertaining to the matter at hand. Erickson's work produced data (her notes related to various documents), but the underlying data did not have an electronic representation, and hence the resulting database was effectively disconnected from the sources. In many cases, the sources are in electronic form and additional data, resulting from analysis or from other sources, can be related to it. This type of relationship can be conceived of as a set of annotations. This approach is not widespread among historians, but there are some instances of such practices. For example, Mugelli et al. defined an annotation logic in order to isolate and categorize references to sacrifice in Greek tragedy[41]. The underlying corpus, in this case, is the Greek text in electronic form, and the annotations inserted within the text contain all the elements contributed by the historians: the specific location where a sacrifice is identified, the categorizing of the sacrifice, or the disambiguation of the discourse about the ritual from the ritual itself, for instance. Processing the annotated text can then result in this case in a database of ritual events. Barker et al. used a very similar framework in order to study spatial references in Herodotus[42]. They relied on the English translation of the Greek text, and after having cleaned the data from Perseus[43] they added annotations tracking the geographical information relevant to them; a copy of these annotations with the surrounding text were then stored in a database.

In a recent methodological paper, Hoekstra and Koolen generalized these notions of organized data for historical research through the use of *data scopes*, which they defined as "the process through which different views on research data are created that are relevant to a specific research question[44]". In their perspective, a corpus for historical research should be created with a general

[40]See Erickson (2013).

[41]See Mugelli et al. (2017).

[42]See Barker et al. (2013).

[43]Crane (2012).

[44]Hoekstra and Koolen (2019), p. 80.

data structure in mind, where different related sources of information can be easily paired in order to better gain new insights. In effect, this corresponds to transforming a general corpus (from whatever source may be available) into an oriented corpus. In their analysis of the interpretation of data by historians, Gibbs and Owens did not focus on a particular way in which data should be represented, but concentrated on the general idea of sharing data processing methodology, which can be understood as a more general perspective than data scoping. They nevertheless lamented the fact that the manner in which data is used by historians is rarely well documented and made accessible[45].

In historiography, the creation of a digital corpus appears as a complex exercise for which, to the dismay of many of the researchers whose work we mentioned above, there is no clear and unique methodology. Further, the amount of detail provided to readers of historical research, when such work is carried out, varies greatly. We may wonder if there is, on the other hand, greater methodological unity in how such corpora are exploited.

Exploiting Corpora with Computational Linguistics

The large-scale digitization of historical material does not systematically require the design of complex databases; the effort sometimes mainly consists of ensuring the quality of the resulting electronic text, and in the storage of all relevant metadata pertaining to the original documents. In classics, it is possible that the relative paucity of textual sources, by making the processing of the entire corpus a reachable objective, encouraged digitization efforts. As Barker and Terras pointed out, it has become quite common in Ancient history to rely on detailed text searches, and there are a variety of dedicated tools for this purpose[46].

Some use cases of text corpora involve more advanced computational linguistics. One example is the medieval charters text corpus described by Magnani, which gathers structured but heterogeneous texts[47]. Such an electronic corpus, as is typical, is designed to be "simple, multiple, open and free of access[48]". There are countless similar electronic corpora that gather historical sources in such a fashion. This particular case, however, has given rise to advanced quantitative meta-analyzes of the texts, using the full range of text processing tools[49].

[45] "Despite some recent methodological experimentation with data, historians have not been nearly as innovative in terms of writing about how they use it", see Gibbs and Owens (2013), p. 163.

[46] See Barker and Terras (2016).

[47] See Magnani (2017b).

[48] Magnani (2017b), p. 64.

[49] See Perreaux (2021).

Indeed, another important use of digital tools includes the automatized distant reading of very large volumes of text. The quantitative analysis of literary texts is not recent, and it has reached a certain maturity[50]. The systematic exploitation of textual data has now become widespread and standard enough that there are programming manuals focused on this type of exercise, covering all the standard operations one might need. The techniques discussed in specialized publications only fifteen years ago[51] made use of series of adhoc tools, which demanded that users manipulate a variety of targeted methods. The tools now available allow their users to program all forms of lexical and semantic analyzes within a unified framework, in a process conceived of as a *pipeline*, comparable to the data analytics suites used in the hard sciences[52]. These various techniques are generally geared towards helping users analyze a given corpus, but they are not specifically aimed at the design of a corpus. Further, they need in some cases a pre-processed corpus, and may not be able to operate on raw text. There are still certain domains for which a very specific technical expertise is required, such as the analysis of the networks stemming from literary texts[53], or the analysis of intertextuality through the methods of computational biology[54]. Specialized computational linguistics domains such as logometrics or sentiment analysis have made their way into historical practice. For example, in contemporary French history, Magali Guaresi applied factor analysis on a large corpus of speeches by members of the French Parliament in order to characterize the evolution of these discourses[55]. Recently, classicists and computer science specialists have collaborated on an empirical study of the quality of automatized sentiment analysis in Greek tragedy, and shown that automatic processing yielded good results compared with humans[56].

For McGillivray et al., collecting and processing historical material with computational methods "would be a science if we could learn to automate it[57]". In their view, historians should clearly delineate between what they define as "evidence" and what they define as "claims", so that in a positivist perspective one may separate evidence-based findings from other statements, thanks to the systematic analysis of evidence. Example-based analysis would be, in their view, kept

[50]See, for example, Hoover (2013).

[51]See for example the general description laid out in Burrows (2007).

[52]See for example Jockers and Thalken (2020). In the case of classics, see Burns (2019) who describe a series of useful operations on classical texts; for a more general treatment of pipelines from the standpoint of data analytics, see Wickham and Grolemund (2017), p. 261-268 in particular.

[53]These analyzes resort to various technical domains that are not part of traditional linguistics, see Kenna, MacCarron, and MacCarron (2017).

[54]See for example, in the case of the Latin language, Chaudhuri and Dexter (2017).

[55]See Guaresi (2019).

[56]See Yeruva et al. (2020).

[57]See McGillivray, Wilson, and Blanke (2019), p. 53.

separate from quantitative evidence[58]. Making processes automatic may not, *per se*, serve a fundamental purpose, but it would have the advantage of making the creation of a corpus and some of its processing reproducible, and hence open it to critique. This would recoup with the objective laid out by Gibbs and Owens, by inherently providing detailed documentation on the use and processing of a corpus.

It appears that, while various tools and technical approaches have converged, thanks to general technical progress, there is no centralized perspective about how textual corpora may be processed. The unifying element is the fact that these texts in electronic form always need some amount of processing. Then, although the notion of a corpus in historiographical work, stemming from a long history, is well understood by historians, the conjunction of that notion with the massive expanse of digital methods has not resulted in a clear and unified epistemological framework, and neither in a well defined methodology. The right framework may necessarily need to recognize the fact that there can be no pre-determined framework.

The Corpus as a Process, not a Result

With so much dispersion among approaches, how can we propose a unifying framework? In order to address the issue, we first need to draw a distinction between data and operations. Relying on this distinction, we will argue that corpora need to be conceived of as operations, not as data. Then, we will discuss an example in ancient Greek history.

Data vs. Operations

Data, as the etymology tells us, is what is given, and cannot be worked out otherwise. Operations are applied to data in order to transform (or, etymologically, work) it into something more usable or practical. Operations orientate the data. Hence, data is information that cannot be derived from other information using logical operations, and requires a human input. In computer science, this distinction between data and operations is better known as that between data and algorithm. Since each needs the other to exist, they are naturally intimately related, but are fundamentally and conceptually different[59]. Note that a database, or an ontology, is not simply data, because it combines data and algorithms that describe the manner in which the data may be exploited. Any database can be represented as a combination of raw structured data[60] and algorithms that

[58] See the chart in McGillivray, Wilson, and Blanke (2019), p. 55.

[59] See Wirth (1976).

[60] Typically as tables of numerical or categorical data. Any multi-dimensional dataset can naturally be represented as a two-dimensional array.

define potential relationships between the columns in these tables. Seen as a whole, the database may obfuscate the interactions between data and operations, and makes it seem as though it only contained the resulting dataset, organized as its designer intended.

All the examples of digital corpora we have discussed so far effectively are combinations of data and operations: they never were the straight input as provided by a person. These corpora were nevertheless considered as *things*, as data, focusing on the result of a series of operations rather than on the operations themselves. The precise descriptions of these results, without explaining the series of operations, were therefore incomplete and superficial. If we simply realize that a corpus is not data, but rather a set of operations, it instantly solves all the issues we have noted so far. In addition, an operation could at the limit effectively contain any dataset: an algorithm may simply possess the list of values that constitute the data. In this sense, a process is superior to data, it is denser in informational content.

We have mentioned the notion of a *pipeline* earlier, in the context of modern data analytics, or in computational linguistics. In such a pipeline, the raw input at each stage is transformed in place, and serves as the input for the next stage. From this perspective, an oriented corpus should be seen as a pipeline, that is a series of operations, not as a dataset. The definition of this corpus is the computer code that builds it, whatever the language in which it may be expressed. This code can be analyzed, and it can be run in whole or in part by anyone. Purely seen as a dataset, a corpus cannot be properly analyzed from an external perspective, but once its construction is made entirely explicit, then this process can be fully subject to critique. In practice, historians often follow such a pipeline: for example, obtain some data (in a spreadsheet), transform it, save the clean version, compute some aggregates. However only the end result survives, and the details of all the steps are lost to everyone else. If this entire process is coded as a pipeline or as an algorithm more generally, every single assumption, explicit or not, becomes visible.

Hence, once defined as computer code, or as a pipeline, the corpus contains and makes explicit all the decisions, small and large, made by the historian in cleaning, filtering, completing, or arranging the raw information, in the most concise manner possible: the corpus's definition is the process, and reciprocally. By construction, this perspective on the corpus as *anima*, that is the process which orientates it, also directly expresses the dialog between the research question and the set up of the corpus. Modern data analysis and text edition tools have largely converged, so that there is not such a strong distinction between the two anymore: the text a researcher produces (the historian's annotation work in Rygiel's words) and the computer code that gathers

and processes data exist in the same document. Considering the data organization, processing, analysis and write-up of a historical corpus as a continuum in a seamless process naturally leads to reproducible research[61].

We can relate this view of the corpus back to the *Annales*' co-founder's of the definition of history: for Lucien Febvre, history may be viewed not as a science, but as the scientifically elaborated narrative of the activities and creations of humankind[62]. The idea of carrying out historical research not as a science, but with a scientific aspiration, is clearly furthered by making the historiographical process more replicable, at least for the part that concerns the set up of a corpus.

An Example with Ancient Greek History

In order to illustrate this logic, let us consider an example with ancient Greek theater, where we would want to study the occurrences of decisions. This would require the creation of a corpus of these decisions, so that they can be closely read and examined. The traditional way of tackling this project could benefit from the fact that these texts are available in electronic form: one could easily copy and paste all the Greek text of interest, maybe accompanied with translation, into a word processing software. For each excerpt, one could give a categorization for the decision at hand (whether it is an acceptance or refusal, for example). If instead we followed a more digital-based logic, comparable to the one described by Mugelli et al.[63], we would first create a copy of the Greek text (from Perseus, for example) in some standard TEI format. Then, inside this text, we would add markers that designate decisions, with some specific coding logic to distinguish between different sorts of decisions. We may also correct the text, if there are issues in how it has been established. Processing the annotated text could then result into a database, containing all the text excerpts with their categorization, amounting to a much more structured and easily exploitable corpus than a list of quotes in a word processor. This later approach, nevertheless, by creating a new object, severs the link between the original textual source and the resulting corpus.

A process-driven, rather than object-driven, perspective on this corpus would consider the text

[61]The RMarkdown language, for example, combines the data and statistical modeling infrastructure of the R language with the edition capabilities of the Markdown syntax and LaTeX system; it has been suggested as a good framework for reproducible research, see Calero Valdez (2020).

[62]"L'étude, scientifiquement conduite, des diverses activités et des diverses créations des hommes d'autrefois, saisis à leur date, dans le cadre des sociétés extrêmement variées et cependant comparables les unes aux autres (c'est le postulat de la sociologie)", and further "je qualifie l'histoire d'étude scientifiquement menée, et non pas de science", see Febvre (1992), p. 19.

[63]Mugelli et al. (2017).

source as an input that should not be changed in place or copied. If corrections are needed on the text, they are made explicit in the pipeline, as a set of overrides. The historian's work in defining the corpus would be embedded in a separate dataset, simply containing the decisions' location identification (in the form of line and word number, for example) and their categorization. From the text input and the categorization data, by merging the two, one can automatically generate the equivalent of the annotated text if needed. Any textual analysis on the text of the oriented corpus, the decisions themselves, can be easily compared with the same analysis applied to the rest of the text, excluding decisions. In addition, if the Greek text's edition is improved and some words are corrected, then the decision corpus immediately benefits from these improvements. Disambiguating the universal corpus in which one seeks excerpts or categorizations, from the added information produced by the historian, we open the entire process to a much better understanding and to document critique. The overriding principle is that any historian input related to the constitution of a corpus, whether in terms of raw data or processing logic, should be reflected as a step in an algorithm or pipeline, and not as a physical operation, such as clicking on some instruction on a piece of software.

Conclusion

In spite of the essential role played by document corpora in the writing of history, and despite the availability at a large scale of electronic sources and processing capacities, historiography has not truly taken stock of the fundamental change brought about by their combination. We have seen that many researchers complained that historians did not fully embrace the digital, and did not fully exploit the tools at their disposal. This is, to some extent, missing the forest for the tree. The issue is not whether one should use such keyword search on some database rather than spend a few hours with a large dictionary. The issue is that the corpus has been considered as a body, as data, rather than what it has fundamentally become, a process.

Once this distinction is made, and once historians realize that this corpus is dead, they will be able to fully attain the scientific aspiration that Febvre had in mind. In particular, many research journals focused on the more quantitative aspects of history nowadays ask that contributors provide the underlying data, if any, supporting their analyzes. This requirement would be more beneficial to all if it was instead phrased as a request for the code or the algorithms that supported the analysis.

References

Barker, Elton, Leif Isaksen, Nick Rabinowitz, Stefan Bouzarovski, and Chris Pelling. 2013. "On Using Digital Resources for the Study of an Ancient Text: The Case of Herodotus' *Histories*." In *The Digital Classicist 2013*, edited by Stuart E. Dunn and Simon Mahony, 45–62. Bulletin of the Institute of the Classical Studies 122. London: University of London Press.

Barker, Elton, and Melissa Terras. 2016. "Greek Literature, the Digital Humanities, and the Shifting Technologies of Reading." *Oxford Handbooks Online* 2016: 1–25.

Bermejo Barrera, Jose Carlos. 2001. "Making History, Talking About History." *History and Theory* 40 (2): 190–205.

Boutier, Jean. 2014. "L'usage historien des archives." In *Corpus, sources et archives*, by Jean Boutier, Jean-Louis Fabiani, and Jean-Pierre Olivier de Sardan, 9–22. Études et travaux de l'IRMC. Tunis: Institut de recherche sur le Maghreb contemporain.

Burns, John, Alan Brenner, Keith Kiser, Michael Krot, Clare Llewellyn, and Ronald Snyder. 2009. "JSTOR - Data for Research." In *Research and Advanced Technology for Digital Libraries*, edited by Maristella Agosti, José Borbinha, Sarantos Kapidakis, Christos Papatheodorou, and Giannis Tsakonas, 416–419. Lecture Notes in Computer Science. Berlin, Heidelberg: Springer.

Burns, Patrick J. 2019. "Building a Text Analysis Pipeline for Classical Languages." In *Digital Classical Philology: Ancient Greek and Latin in the Digital Revolution*, edited by Monica Berti, 159–176. Age of Access? Grundfragen der Informationsgesellschaft 10. Berlin ; Boston: De Gruyter Saur.

Burrows, John. 2007. "Textual Analysis." In *A Companion to Digital Humanities*, edited by Susan Schreibman, Ray Siemens, and John Unsworth, 323–347. Blackwell Companions to Literature and Culture 26. Malden, MA: Wiley Blackwell.

Calero Valdez, André. 2020. "Making Reproducible Research Simple Using RMarkdown and the OSF." In *Social Computing and Social Media. Design, Ethics, User Behavior, and Social Network Analysis*, edited by Gabriele Meiselwitz, 27–44. Lecture Notes in Computer Science 12194. Cham: Springer International Publishing.

Camar, Françoise, Véronique Samouiloff, Renaud Dalmar, Séverine Cassar, and Anne Fleury. 2017. *Arlette Farge, une historienne des vies ordinaires dans ses archives. La*

fabrique de l'histoire. Paris: France Culture. https://www.franceculture.fr/emissions/la-fabrique-de-lhistoire/historiennes-22-arlette-farge-une-historienne-des-vies-ordinaires-dans-ses-archives.

Chaudhuri, Pramit, and Joseph P. Dexter. 2017. "Bioinformatics and Classical Literary Study." *Journal of Data Mining and Digital Humanities* Special Issue on Computer-Aided Processing of Intertextuality in Ancient Languages: 1–8.

Crane, Gregory R. 2012. "Perseus Digital Library." April 2012. http://www.perseus.tufts.edu.

Erickson, Ansley T. 2013. "Historical Research and the Problem of Categories: Reflections on 10,000 Digital Note Cards." In *Writing History in the Digital Age*, edited by Jack Dougherty and Kristen Nawrotzki, 133–145. Digital Humanities. Ann Arbor: University of Michigan Press.

Febvre, Lucien. 1934. "Comptabilité et Chambre des Comptes." *Annales d'histoire économique et sociale* 6 (26): 148–153.

———. 1992. *Combats pour l'histoire*. Originally published in 1952. L'ancien et le nouveau 12. Paris: Armand Colin.

Foucault, Michel. 1969. *L'archéologie du savoir*. Paris: Gallimard.

Gibbs, Fred, and Trevor Owens. 2013. "The Hermeneutics of Data and Historical Writing." In *Writing History in the Digital Age*, edited by Jack Dougherty and Kristen Nawrotzki, 159–170. Digital Humanities. Ann Arbor: University of Michigan Press.

Guaresi, Magali. 2019. "La logométrie en histoire : une herméneutique numérique. Exploration d'un corpus de professions de foi électorales de député-e-s (1958–2007)." *Digital Studies/Le champ numérique* 9 (1): 14.

Hitchcock, Tim. 2013. "Confronting the Digital: Or How Academic History Writing Lost the Plot." *Cultural and Social History* 10 (1): 9–23.

Hoekstra, Rik, and Marijn Koolen. 2019. "Data Scopes for Digital History Research." *Historical Methods: A Journal of Quantitative and Interdisciplinary History* 52 (2): 79–94.

Hoover, David L. 2013. "Quantitative Analysis and Literary Studies." In *A Companion to Digital Literary Studies*, edited by Ray Siemens and Susan Schreibman, 517–533. Blackwell Companions to Literature and Culture. Malden, MA: Wiley Blackwell.

Jockers, Matthew L., and Rosamond Thalken. 2020. *Text Analysis with R: For Students of Literature*. Quantitative Methods in the Humanities and Social Sciences. Cham: Springer International Publishing.

Kenna, Ralph, Máirín MacCarron, and Pádraig MacCarron, eds. 2017. *Maths Meets Myths: Quantitative Approaches to Ancient Narratives*. Understanding Complex Systems. Cham: Springer International Publishing.

Magnani, Eliana. 2017a. "Qu'est-ce qu'un corpus ? Compte-rendu de la journée d'études." *Carnets de l'IHRT Chartae Burgundiae Medii Aevi*: online.

———. 2017b. "Un corpus structuré et hétérogène de textes latins médiévaux (Bourgogne, V^{e}-XV^{e} siècle)." *Bulletin du CERCOR* 41: 59–65.

Mayaffre, Damon. 2002. "Les corpus réflexifs : entre architextualité et hypertextualité." *Corpus* 1: 1–14.

McGillivray, Barbara, Jon Wilson, and Tobias Blanke. 2019. "Towards a Quantitative Research Framework for Historical Disciplines." In *Proceedings of the Workshop on Computational Methods in the Humanities 2018*, 2314:53–58. Aachen: CEUR Workshop Proceedings.

Mugelli, Gloria, Andrea Bellandi, Federico Boschetti, and Fahad Khan. 2017. "Designing an Ontology for the Study of Ritual in Ancient Greek Tragedy." In *Proceedings of Language, Ontology, Terminology and Knowledge Structures Workshop*, edited by Francesca Frontini, Larisa Simeunović, Špela Vintar, Fahad Khan, and Artemis Parvisi, 96–105. Montpellier: Association for Computational Linguistics.

Nawrotzki, Kristen, and Jack Dougherty. 2013. "Introduction." In *Writing History in the Digital Age*, edited by Kristen Nawrotzki and Jack Dougherty, 1–18. Digital Humanities. Ann Arbor: University of Michigan Press.

Offenstadt, Nicolas. 2011. "Archives, documents, sources." In *Historiographies : concepts et débats*, edited by Christian Delacroix, François Dosse, Patrick Garcia, and Nicolas Offenstadt, 1:220–231. Folio Histoire 179. Paris: Gallimard.

Perreaux, Nicolas. 2021. "Possibilities, Challenges and Limits of a European Charters Corpus (*Cartae Europae Medii Aevi* - CEMA)." hal-03203029. Paris: HAL archives ouvertes. https://hal.archives-ouvertes.fr/hal-03203029.

Pflug, Gunther. 1971. "The Development of Historical Method in the Eighteenth Century [1954]." *History and Theory* 11: 1.

Rygiel, Philippe. 2011. "L'enquête historique à l'ère numérique." *Revue d'histoire moderne et contemporaine* 58-4 bis (5): 30.

Treffort, Cécile. 2014. "Le corpus du chercheur, une quête de l'impossible ? Quelques considérations introductives." *Annales de Janua* 2: online.

Wickham, Hadley, and Garrett Grolemund. 2017. *R for Data Science*. Beijing ; Boston: O'Reilly.

Wirth, Niklaus. 1976. *Algorithms + Data Structures=Programs*. Prentice-Hall Series in Automatic Computation. Englewood Cliffs, NJ: Prentice-Hall.

Yeruva, Vijaya Kumari, Mayanka Chandrashekar, Yugyung Lee, Jeff Rydberg-Cox, Virginia Blanton, and Nathan A Oyler. 2020. "Interpretation of Sentiment Analysis in Aeschylus's Greek Tragedy." In *Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*, 138–146. Barcelona: SIGHUM.