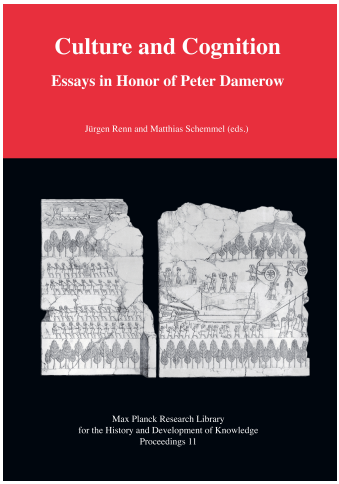
Max Planck Research Library for the History and Development of Knowledge

Proceedings 11

Mark J. Schiefsky:

Beyond Archimedes: The History and Future of the Arboreal Software



In: Jürgen Renn and Matthias Schemmel (eds.): *Culture and Cognition : Essays in Honor of Peter Damerow*
Online version at `http://mprl-series.mpg.de/proceedings/11/`

ISBN 978-3-945561-35-5
First published 2019 by Edition Open Access, Max Planck Institute for the History of Science.

Printed and distributed by:

The Deutsche Nationalbibliothek lists this publication in the Deutsche Nationalbibliografie; detailed bibliographic data are available in the Internet at `http://dnb.d-nb.de`

Chapter 20
Beyond Archimedes: The History and Future of the Arboreal Software

Mark J. Schiefsky

This chapter describes the history and rationale of the development of Arboreal, a software application originally designed in the course of the Archimedes Project to enable the reading and analysis of structured XML documents. It also provides an update on current development of the software and sketches some directions for further work. Peter Damerow inspired and contributed to the development of this software at every stage from its inception until his death, and its existence would be unthinkable without him. I hope that this history will illuminate an important dimension of his work and convey the essence of his compelling vision for the place of information technology in humanistic scholarship. Above all, I hope that my description of the present state of this software and its future development will demonstrate that this vision is still very much alive today.

History and Rationale

The Archimedes Project was initiated by Peter Damerow and Jürgen Renn at the Max Planck Institute for the History of Science (MPIWG) in the late 1990s. It was conceived as the digital component of a major research project of Department I of the Institute on the long-term development of mechanical knowledge from the ancient world up to the early modern period. The goal of the project was to exploit the potential of emerging digital methods to study the content of mechanical knowledge and to disseminate the results of this research. Earlier work at the Perseus Project of Tufts University had demonstrated the great power of new tools for linguistic analysis, such as morphological analyzers and digitized dictionaries, for the creation of a new kind of online environment for the reading of ancient Greek and Latin texts. The possibility of extending this approach to the history of science was an exciting one, and very much called for by the large volume of textual sources involved in the study of a discipline like mechanics. More specifically, Archimedes was motivated by the need to represent the conceptual structure of mechanical thinking—relations between concepts such as

"force," "weight," and "motion," such as the notion that "motion implies force" that is ubiquitous in early physical thought. To achieve a digital representation of the conceptual content of mechanical texts at this level of detail was a challenging task that went well beyond anything then existing in the Internet.

In 1999, I came to the MPIWG as a postdoc to work on the Archimedes Project, with a mandate to introduce the technology developed by the Perseus Project to the research environment of the Institute. This was a year of intellectual joys for me, not least because I had the remarkable privilege of working closely with Peter Damerow on a daily basis. We discussed everything from the origin of writing and mathematics in the third millennium BCE to markup practices for electronic texts and the history and culture of contemporary Berlin. My own scholarly perspective as a classical philologist and historian of science was immensely enriched, in ways that have continued to shape my intellectual outlook to this day. One topic that was of constant concern in our daily work during this period was the design of electronic working environments that could meet the demands of the Archimedes Project for the detailed representation of the content of mechanical knowledge. I quickly became familiar with the wide range of ingenious prototypes that Peter had created for the Macintosh using FileMaker. These were based on a simple idea: a text was split up into sentences which were then loaded into a FileMaker database, in which each sentence was an individual record. This strategy made it possible to browse through the text sentence-by-sentence, to track the terminology in which ideas were expressed in individual sentences, and to add translations or comments on sentences in a systematic manner. FileMaker's powerful indexing and viewing functions made it possible to view clusters of sentences together and to see connections between the data that were not otherwise apparent. And all of this could be achieved without any knowledge of formal programming languages. Even though the technological basis was quirky and idiosyncratic, it was clear that a powerful vision lay behind it. The use of technology was motivated by scholarly questions and tailored to scholarly aims. Peter consistently emphasized the importance of dynamic interactivity of the computing environment and human scholarly input, and the need to keep any technological solution as simple as possible. And of course the overall goal was to create resources that would be freely available online without any restrictions on access due to copyright considerations.

During the year that I spent at the MPIWG we succeeded in incorporating the results of the Perseus morphological analysis software and online dictionaries into these working environments. But despite the ingenuity of Peter's FileMaker prototypes, it was clear to all of us involved in the project that Archimedes demanded more powerful and robust tools. One of Peter's early decisions had been that the project texts would be tagged using the XML markup language, which

was still relatively new at that time. This was exactly the right call, given that XML is now the de facto standard for text markup; at the time, however, there was very little user-friendly software available for working on XML texts, so the decision to use XML was something of a leap into the unknown. By the end of 1999, when I returned to Harvard as an assistant professor of Classics, we had in hand the first set of digitized texts making up the Archimedes corpus: a collection of early modern writings on mechanics in Latin and Italian by authors such as Guidobaldo del Monte, Tartaglia, and Galileo. Correcting and tagging these in a simple XML format took up the lion's share of the project's efforts in the subsequent year. In 2000 we also secured three years of funding for Archimedes from the Deutsche Forschungsgemeinschaft and the National Science Foundation; in 2001 Malcolm Hyman, a brilliant young linguist and Classicist, joined the project as a postdoc. At this point Malcom convinced me that it was time to bite the bullet and create new software from scratch that would address the basic technical challenges of making it possible to work with XML texts in the ways required by the Archimedes Project. Thus Arboreal came to be.

I will now describe the basic features of this software in the form that it reached during the Archimedes Project; for convenience we may think of this as "Arboreal 1.0."[1] Arboreal is conceived on the analogy of a traditional web browser, albeit one that works on XML rather than HTML texts. But one immediate difference is that when the user opens an XML document in Arboreal, instead of a single window we see two panes, one depicting the XML tree structure on the left and another content pane on the right. The user can navigate through the document by selecting nodes in the tree pane. As elements in the tree are selected in the tree pane they are rendered in the content pane. Arboreal allows for very complex XML structures (e.g. a text with many deletions and supplements indicated in the markup) to be rendered in whatever way is suitable for the application in question; the display of XML tags can be toggled on or off at will. Thus the XML markup is brought to life in a way that makes it accessible to the user, while the underlying format of the document is unchanged. Clicking on any word in the content pane reveals a pull-down menu that provides access to morphological and dictionary information. This information is generated on the Harvard Archimedes server (http://archimedes.fas.harvard.edu), on which we have implemented a unified frontend to various backend morphological analyzers (http://archimedes.fas.harvard.edu/donatus). Texts can contain as many languages as are supported by the software on the server, as long as these are tagged in the XML markup, and only a single server request is needed to generate the complete set of morphological analyses for all languages in a document.

[1] In terms of the actual numbering, the last version on which Malcolm Hyman worked was designated "Arboreal 5.16."

Searching can be carried out using lexical forms and regular expressions; in addition, search results can be used to navigate through the document tree. Arboreal allows for the annotation of individual words or groups of words as instances of terms, which can be manipulated and visualized in a special term editor. Terminology annotations can be saved as XML documents and subjected to further analysis. Arboreal also allows for the study of different texts in parallel to one another, making it possible to carry out systematic comparisons between texts and translations, and to create and edit translations. Finally, Arboreal enables the creation of XML content of various types. Arbitrary XSLT scripts can be applied to the text currently loaded in the main window; moreover, the terminology, morphology, and matching files that underlie the program's other functions are also in XML format.[2]

The development of this software resulted from years of intensive effort and collaboration between many individuals with different intellectual approaches and technical competences. It would have been impossible without the programming genius and intellectual vision of Malcolm Hyman, who had the original insight that such software was possible, the ability to bring it into being, and the determination to do so. The initial stage of development took place at Harvard University, in close collaboration with Peter Damerow, Jürgen Renn, and other members of Department I of the MPIWG. There were many trips between Boston and Berlin in those years. When Malcolm took up a position at the Institute in 2005, he was able to work even more closely with the colleagues there. During the period between 2005 and Malcolm's tragic death in 2009, Arboreal was used intensively in creating a translation of a Chinese text on mechanics at the MPIWG and for studying the terminology and deductive structures of Euclid's *Elements* (Schiefsky 2007).

While Arboreal was designed with the needs of the Archimedes Project in mind, it turned out to be a highly general tool. In fact, Arboreal embodies several key features that challenged—and continue to challenge—the basic way in which the Internet functions as a medium for the creation and representation of knowledge. First, Arboreal moves beyond browsing; it enables the creation of richly structured digital content in providing facilities for term annotation and the generation of new XML documents from existing texts. Second, it also moves beyond the search-dominated paradigm of the current Internet. Although Arboreal has powerful capabilities in this regard, searching is not conceived as an end in itself; rather, the program is designed to make search results the starting point for further

[2]For a longer description of the functionality of "Arboreal 1.0," see Schiefsky (2007). This description omits a number of functions provided by Arboreal that are not directly relevant to my argument here. In particular, I should note that the program is also designed to interact smoothly with image repositories and tools for image annotation, to enable the analysis of visual as well as textual content.

analysis. Third, Arboreal provides a robust platform for multilingual computing, and a model for providing the user with integrated access to diverse linguistic resources. Finally, Arboreal provides for a distinctive kind of interactivity, since the files it generates are themselves XML files that can be subjected to further analysis by the software itself. In the most general terms, Arboreal is a tool that contributes to the long-term project of making the Internet into what has been called an "epistemic web"—a domain in which unstructured, unanalyzed data is transformed into structured information and knowledge. In this vision, which descends from Peter Damerow's insights but represents the fruit of the entire course of development described in this history, the Internet becomes a vehicle for transmitting knowledge, understanding, and, we may hope, also wisdom.[3]

Since the formal conclusion of the Archimedes Project in 2004, a great deal has changed in the universe of the Internet. Social networks have exploded, enabling an exponential increase in user-generated content. Google Translate now provides a crude translation for all the world's main languages and offers the eventual possibility of a linguistically transparent Web. Natural language processing applications have greatly advanced beyond the stage of context-free morphological analysis that was still the state of the art at the time that Archimedes began. And there are vastly more digital corpora now available for research. The "big data" approach—using statistical methods to analyze and extract meaning from huge corpora—has enabled entirely new questions to be asked and new approaches to be taken to traditional questions.

Yet in many ways the ideal of the epistemic web seems more remote than ever. In the area of digital scholarship, researchers seem trapped in the "browsing" and "search" paradigms. The massive success of Google has contributed to a tendency to reduce analysis to searchability; the problem of giving the user informed assistance in *what to search for* has been neglected. Social networks offer the user the possibility to create content, but this tends to be limited to unstructured text or images. There is still an urgent need, then, for software that embodies the distinctive features of Arboreal as outlined above. In some ways the current environment renders the need for such software all the more acute. Recent challenges posed by the "big data" movement suggest that statistical approaches to large corpora may make traditional scholarly (and other kinds of) analysis irrelevant. Indeed it has recently been argued that big data and the associated modes of analysis are on the verge of eliminating the need for models and explanatory hypotheses in scholarship and in science (Anderson 2008).

But while there may be many ways of creating knowledge out of unstructured information, there is still a crucial place for human input—both in determining the questions that should be posed to automatic systems and in interpreting the

[3] For the concept of the epistemic web, see Hyman and Renn (2012).

results they produce. "Big data" approaches have their place in the sciences and—increasingly—in the humanities, but it is a mistake to think that they can answer all interesting questions. Informed input is needed not just to interpret results, but also to pose the questions. What is particularly lacking in the arguments in support of the "big data" approach is a sense of the power of interactivity—of the way in which automatically generated results only gain meaning from informed questions, and are shaped by them. Reflecting on tools like Arboreal can help us to see how technology can be used to foster these humanistic ends. Indeed I would argue that the potential of the technology itself will not be fully realized unless such interactivity is kept front and center in the development process.

To return to history. The tragic death of Malcolm Hyman in 2009 dealt a severe blow to Arboreal's development. With Peter Damerow's passing in 2011, we lost another of the program's original sources of inspiration. But I am delighted to announce that development of the code has recently resumed with the assistance of two French colleagues, Professor Said-Esteban Belmehdi of the Université de Lille and his graduate student Julien Razanajao. Working in close collaboration with me, Belmehdi and Razanajao have updated the code to work with contemporary versions of Java and succeeded in integrating some powerful new features.[4] In the remainder of this paper I shall describe these features, give some examples of their use, and outline some of the principal goals that remain for further work.

"Arboreal 2.0": Networks and Visualization

The principal innovation in this new and improved version of Arboreal is the ability to generate networks and to visualize them using the Gephi software library (http://gephi.org). This provides a standard format (.gexf) for encoding graphs in XML, as well as a powerful set of algorithms and rendering tools for viewing and analyzing graph data. The current version of Arboreal makes it possible to generate and render two kinds of graphs: graphs of the distribution of morphological variants of a given word across sections of a text, and semantic networks expressing the relations between different terms. I shall illustrate these with examples drawn from the history of mechanics and the texts of the Archimedes project corpus.

[4] It is a pleasure to acknowledge the programming skills and dedication of my two French collaborators, who worked tirelessly, successfully, and without any special institutional or financial support to decipher and extend Malcolm Hyman's brilliant but very complex Java code. Without their work none of the analysis I describe in the rest of this paper would have been possible, and the future of Arboreal would be in serious doubt.

Morphological Graphs

Morphological graphs, as defined above, have their principal use in the exploration of the relationships between the language and formal structure of texts. Consider the example of the very first ancient Greek text dedicated to theoretical mechanics, the *Problemata Mechanica* or *Mechanical Problems* attributed to Aristotle. This text contains a long introduction on the wondrous properties of the circle and circular motion, which in the author's view underlie the explanation of all mechanical movements. The introduction is followed by a set of 35 "problems" or questions that are posed using a standard formulation, then answered. Thus problem 1 asks why it is that larger balances are (allegedly) more accurate than smaller, and the author goes on to give an explanation of this fact in terms of circular motion. The text states that the balance is explained in terms of the circle, the lever in terms of the balance, and all other mechanical movements in terms of the lever. But in fact the author often appeals to circular motion directly rather than the lever. We can see this by considering the following graph of the occurrences of μοχλός ("lever") throughout the text (fig. 20.1). The large nodes represent the text itself (on the left) and the lemmatized form of the term (on the right); the lemmatized form is joined to its different morphological variants, which are themselves linked to sections (i.e. "problems") in the text. From this graph we can see at once that the term μοχλός has a fairly wide distribution across the text although the problems in which it does *not* occur also stand out clearly (these are the nodes arranged concentrically around the "Problemata Mechanica" root node).

A more complex example is provided by Guidobaldo del Monte's *Mechanicorum Liber* (1577), a key text in early modern mechanics that has a clear formal structure based on the Euclidean model. After a preface on the nature and importance of mechanics, Guidobaldo sets out a number of basic assumptions or postulates and goes on in six main sections to treat of the balance as well as the five "mechanical powers" familiar from Greek antiquity: the lever, the wheel and axle, the pulley, the wedge, and the screw. Each of the six sections is clearly divided into propositions, lemmas, and corollaries, which are tagged in the XML markup (just as the different "problems" are in the Aristotelian text mentioned above). If we consider the graph of the distribution of the term *gravitas* or "heaviness" (fig. 20.2), we find a strikingly high frequency in Proposition IV of Book I, "On the balance" (*De libra*; the thickness of the edge indicates frequency). Upon inspection of the text, we see that this is because of a large number of instances of the phrase "center of gravity" (*centrum gravitatis*) in this proposition. The frequency of this term is due to the fact that in Proposition IV, Guidobaldo is arguing against thinkers who claimed on the basis of the medieval theory of "positional heaviness" (*gravitas secundum situm*) that a balance displaced from

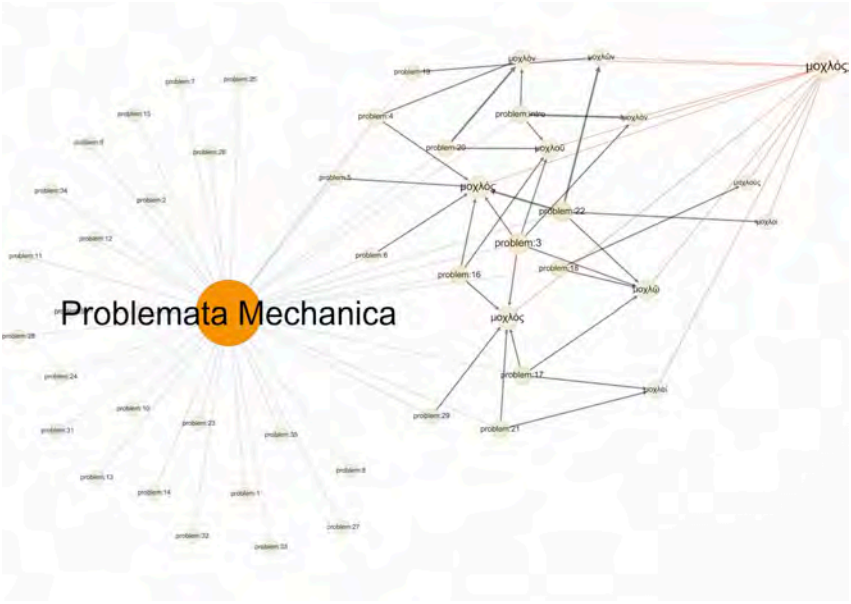


Figure 20.1: Morphological variants of μοχλός ("lever") in the *Problemata Mechanica*.

horizontal equilibrium will return to the horizontal, because the weight that is elevated has more "positional heaviness" than the one that is depressed below the original level. To this medieval theory Guidobaldo opposes the Archimedean theory of center of gravity, arguing that an object suspended from its center of gravity will not move, no matter what its orientation is to the horizontal. This "equilibrium controversy"—as Peter Damerow and Jürgen Renn have called it (Damerow and Renn 2012)—was generated by the tension between two different traditions of mechanical knowledge: one centered on the medieval concept of "positional heaviness" and another rooted in the Archimedean study of centers of gravity. The relevance of the medieval concept to Proposition IV of book I is clearly shown in the morphological graph of the distribution of *situs* (fig. 20.3). Thus, these graphs make it possible to correlate language and formal structure in a precise and visually informative way.

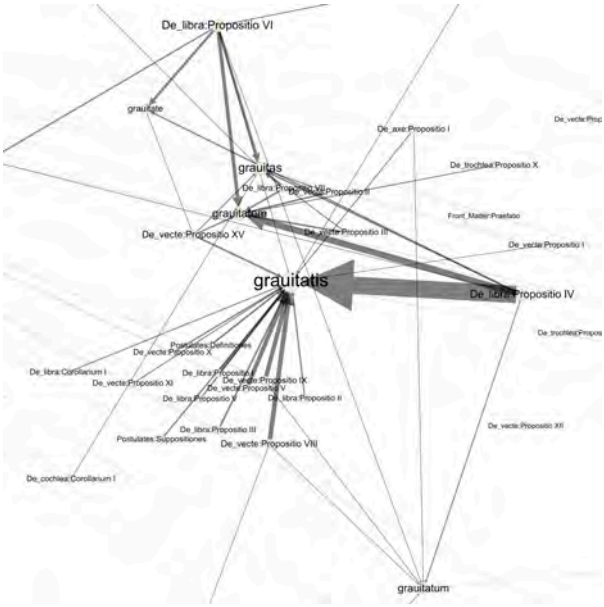


Figure 20.2: Morphological variants of *gravitas* ("heaviness") in Guidobaldo del Monte, *Mechanicorum Liber*.

Semantic Networks

As Malcolm Hyman argued, semantic networks provide a powerful tool for studying conceptual development in the history of science (Hyman 2007). In such networks, individual terms in a text are represented as nodes in a network, and the edges linking them express the strength of their association with other terms. Very crudely speaking, terms that frequently occur in close proximity to one another have a higher association than other terms. Thus, to take an example from Hyman's paper, "force" will have different associations in an article on theoretical physics (e.g. "potential," "field," "electromagnetic") than in an article on the behavior of police towards African Americans in the American South (e.g. "violence," "abuse"). The meaning of a term within a particular text is constituted at least in part by its relations to other terms; "force" in the sense of physical strength exerted by humans is a different concept from "force" in the context of theoretical physics, although the *term* used is the same. A semantic network is thus a kind of model of the conceptual structure of a text, in which terms serve as proxies for concepts.

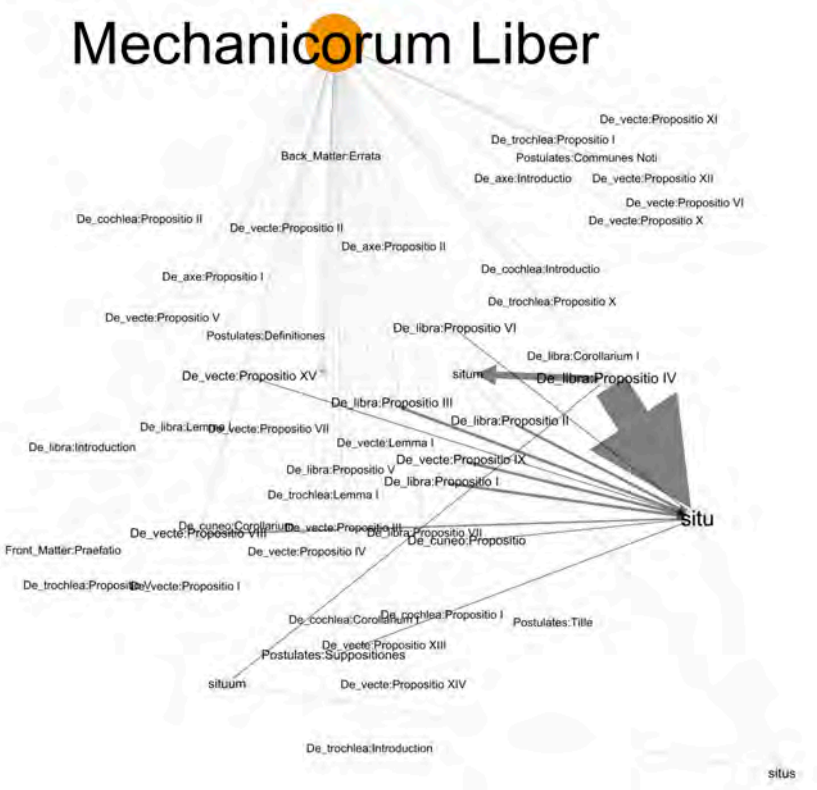


Figure 20.3: Morphological variants of *situs* ("position") in Guidobaldo del Monte, *Mechanicorum Liber*.

Such an approach is particularly useful in studying the semantic field of a concept such as *force* in texts that span a range of centuries and languages. In the early history of mechanics force is typically seen as a cause of motion; but the language of force can also be used of effects rather than causes (e.g. "a forceful blow"). There is also the distinction between physical force and a more generalized power or capacity to affect, as well as that between a force that acts instantaneously and one that persists. The Greek term δύναμις, for example, tends to refer to a general capacity or faculty to affect, rather than something that causes motion in particular; the best translation is often "power," which also expresses the tendency of a δύναμις to persist through the duration of its activity. The term ἰσχύς is

glossed by Bonitz in his *Index Aristotelicus* as both "motive force" (*vis motrix*) and "bodily strength" (*robur corporis*); yet he also notes that it is typically used as a synonym for δύναμις.[5] A third Greek term, βία, has connotations of violence as well as physical strength, and is particularly important for its use to express the distinction between "natural" and "violent" motion in the Aristotelian tradition. Thus, Greek presents us with a set of terms that overlap in their meanings, though each has distinctive nuances. Similar points might be made for the Latin (*potentia*, *virtus*, and *vis*) and Italian (*forza*, *potenza/possanza*, and *virtù*) terminology on the basis of appropriate lexica. While these remarks on general usage are important, understanding the terminology of force in a particular text requires taking account of the way in which the term is actually used. It is this that the semantic network method enables us to investigate in a precise, rigorous and repeatable manner. In this perspective, the meaning of a term in a text is constituted by its place in the semantic space—a web of connections that model the text's conceptual structure.

For the generation of semantic networks the current version of Arboreal implements the Semantic Vectors package released on GitHub (https://github.com/semanticvectors/semanticvectors/wiki). I will simply sketch the general idea here, referring to the online documentation for the details of this particular implementation. The basic idea is to represent the distribution of terms in a document via a term-document matrix. Thus, for example, the rows of the matrix may correspond to segments of text while columns correspond to particular terms; in this simple model, the value of any element of the matrix (r, c) is the number of occurrences of term c in segment r. Once such a matrix is constructed we can apply statistical methods and linear algebra techniques to derive measures for the similarity between different rows (comparing segments to segments) or columns (comparing terms to terms). A key step is dimensionality reduction, in which transformations such as singular value decomposition are used to reduce the size of the matrix; the reduced matrix is then interpreted as a representation of the document in semantic space. The net effect of this is to eliminate the "noise" caused by phenomena such as synonymy: if terms A and B are both found regularly in conjunction with the same cluster of terms C, A and B will end up very close to one another in the semantic space. Another way of looking at this is that A and B express the same *concept* within the semantic space, which is thus a model of the text's conceptual structure. Once the semantic space has been constructed, we determine the association of terms to one another using standard metrics such as cosine similarity. These associations become the labels of the edges linking different nodes in the graph.

In analyzing a particular text, we begin by reducing morphological variants to lexical forms (thus English *is* and *was* → *be*, Latin *vires* and *vi* → *vis*) using the

[5] See Bonitz s.v. ἰσχύς, available at http://archimedes.fas.harvard.edu/pollux.

web services of the Harvard Archimedes server (http://archimedes.fas.harvard.edu) or other methods.[6] We then use Arboreal to build the semantic vectors for this reduced document and perform a pairwise comparison to determine the association that each term has with every other. (This step can take a significant amount of time for larger documents. It is also possible to compare a subset of terms T with all the terms N in the document.) The user is given the option to specify parameters used by the Semantic Vectors package, including especially the length of the segments into which the text is divided (shorter segments will result in fewer associates). Once the graph has been generated, the Gephi package provides many different algorithms for rendering the graph data in perspicuous form; these can be performed in any sequence that the user desires. Additionally, threshold values can be specified to restrict the range of associations that are rendered; we find, for example, that rendering all edges that have a score of 0.65 or above is sufficient both to eliminate noise and to reveal interesting structural features. After the graph is rendered the user can easily select subgraphs by clicking on particular nodes. Arboreal thus provides a highly interactive implementation of the Gephi algorithms that is in some ways more powerful than the Gephi application itself. Finally, I note that Arboreal can export graphs as XML files (using the .gexf format) which can then be analyzed using a range of graph-theoretic analytical tools.

Let me now turn to some examples drawn from the history of mechanics. If we apply this method to the *Problemata Mechanica*, the result is a set of groups divided into an outer ring and inner clusters (see fig. 20.4).

At the center of the largest cluster is found the term κινέω, "to move" and its close associates such as βάρος "weight," ῥᾴδιος "easy/ easily," and κίνησις "motion" (see fig. 20.5).

In this graph the larger nodes have higher *degree*, where the degree of a node is defined as the number of edges connecting it to other nodes. The rendering algorithm has driven nodes of high degree to the center. Indeed, the terms that appear as central here are the terms with the highest degree of all nodes in the graph: κινέω has degree 64, βάρος 50, ῥᾴδιος 42, and κίνησις 41, where the average degree is 3.195.[7] Now κινέω is of course a key term in the *Problemata,* which is very much concerned with the issue of moving weights (βάρος) by the use of a force (δύναμις or ἰσχύς). Indeed the author in the introduction almost immediately raises the general question why small forces can move great weights (implying that the normal or natural course of events is for a force to move a

[6] Among the most useful other tools are the Tree Tagger developed by Helmut Schmid at the Institute for Computational Linguistics of the University of Stuttgart (http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/), which provides lemmatization and part-of-speech tagging for Latin, Italian, and other languages in the Archimedes corpus.

[7] Results according to the Gephi application.

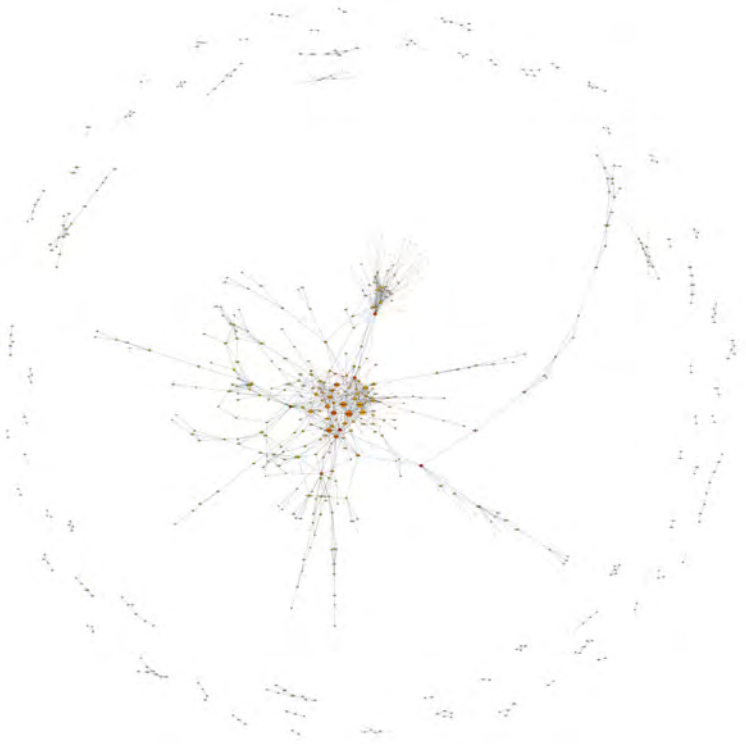


Figure 20.4: Semantic network of the *Problemata Mechanica*.

weight that is equal to it). Moreover, the text's fundamental explanatory principle involves circular motion, and in particular the fact that the movement of a point farther from the center of a circle is quicker than one that is closer to it, assuming the two points lie along the same radius. Hence the presence of κύκλος "circle" and μέγας/μικρός "large" and "small" in this graph. Clearly the semantic analysis is capturing essential aspects of the text's conceptual content; nodes with high degree correspond to terms that are especially significant in some way. With some knowledge of the content of the text, we can explain this significance and supply meaning to the edges that goes beyond a simple numerical score.

We can study the associates of κινέω in more detail by selecting the node and rendering its associates; for this purpose it is convenient to set the lower limit of edge strength to 0 to bring out all the associations (fig. 20.6). We find among

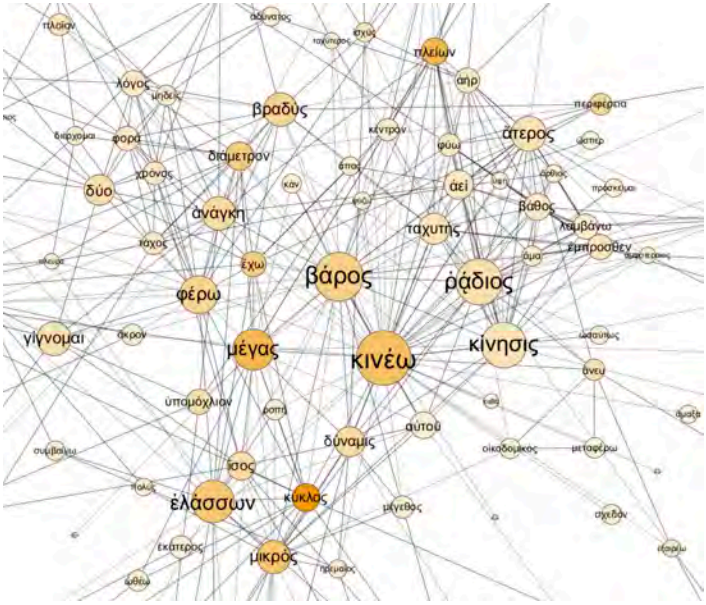


Figure 20.5: Semantic network of the *Problemata Mechanica* (central section).

the associates of κινέω not only ἰσχύς and δύναμις but also the term ῥοπή, which refers to the "inclination" or "swing" of the balance and to something (a small weight) that produces that inclination. We see that μέγας ("large") is associated with terms for force—δύναμις and ῥοπή—as well as βάρος ("weight"). For this author, mechanical movements are generated and explained by the quantitative relationships between force and weight.

The *Problemata* was a seminal text in early modern mechanics thanks to works such as the 1525 Latin translation by Niccolò Leonico Tomeo and the 1547 Latin *Paraphrasis* by Alessandro Piccolomini.[8] Examining the central clusters of the semantic networks of these texts reveals the same basic picture that we find in the Greek, with *moveo/muovere* "to move" at the center with a close linkage to *circulus/circulo* "circle" and *pondus/peso* "weight" (fig. 20.7; fig. 20.8). The graph suggests that Tomeo uses *potentia* to denote the force that causes the weight to move; in Piccolomini, however, the corresponding term is *forza*. The similarity of the place of these terms in their respective graphs suggests that the

[8]For the present analysis I have used the 1582 Italian translation of the latter work. See Drake and Rose (1971) for the basic bibliographical and biographical information pertinent to this literature.

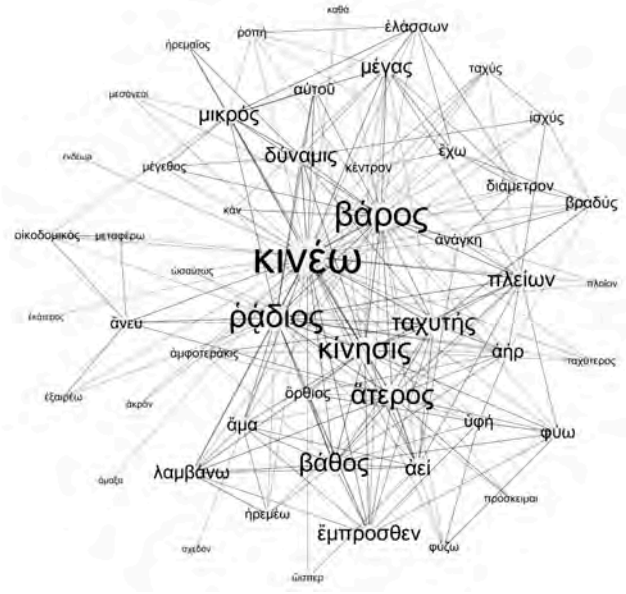


Figure 20.6: Semantic network of κινέω ("to move") in the *Problemata Mechanica*.

terminological difference is simply a matter of stylistic preference, and that the two terms express the same concept. Closer analysis of the texts is needed to confirm this hypothesis, but it is a strength of the present method that it points the reader toward such investigation.

In contrast to the concern with movement in the *Problemata* tradition, an important strand in sixteenth-century mechanics focused on problems of static equilibrium and the determination of center of gravity rather than motion. A key example is again Guidobaldo's *Mechanicorum Liber*.[9] Figure 20.9 shows the central section of the semantic network for this text.

The focus on motion has been replaced by an extremely close association between *potentia* ("power"), *pondus* ("weight"), and *sustineo* ("sustain"). Guidobaldo's text is concerned with determining the power (*potentia*) that will sustain a weight—as one weight balances another on the balance. Of course Guidobaldo also writes of moving weights; indeed he echoes the traditional formulation of the basic problem of mechanics dating back to the ancient Greco-Roman world: "to move a given weight by a given power." But his

[9] See Damerow and Renn (2010).

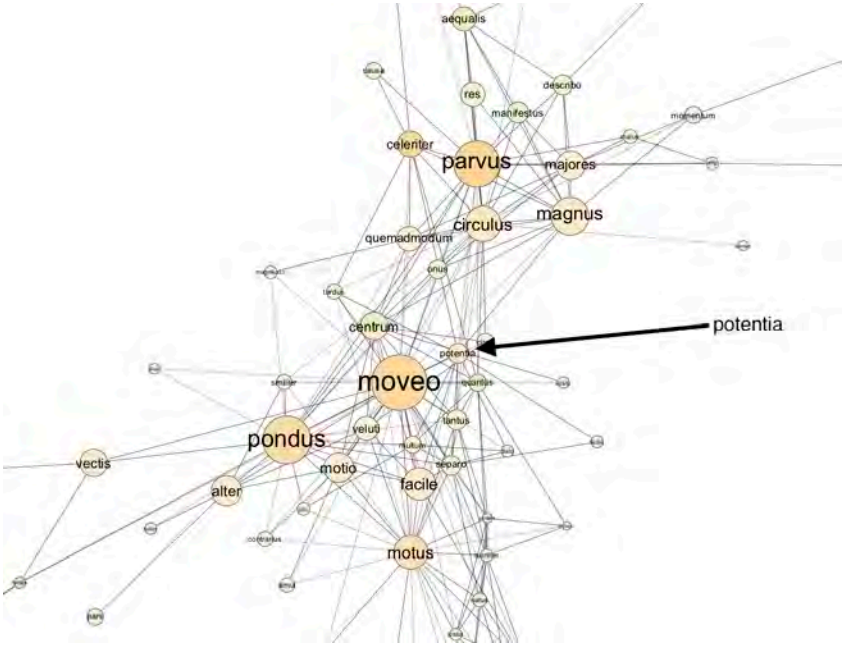


Figure 20.7: Semantic network of Tomeo's Latin translation of the *Problemata Mechanica* (central section).

approach is to determine the power that sustains the weight, on the assumption that movement will ensue if that power is increased. For him, *potentia* denotes a "force" that corresponds to "weight," and does not involve considerations of movement.

The following remark by Filippo Pigafetta from his 1581 Italian translation of Guidobaldo's text brings out the latter's concern with the "sustaining power." It follows proposition 5 in the book on the pulley:

> In this proposition it is shown reasonably that, for two pulleys and one rope, the force (*forza*) will be one-third of the weight […]. Somebody might consider this very dubious, because the pulleys and their attachments, the ropes, and so on offer resistance to the force (*forza*), and also have weight of their own, so that the [calculated] force may not be able to sustain (*sostenere*) the weight. We reply that these things may well offer resistance to the moving of the weight, but not to the sustaining of it; and it is necessary to note carefully that the au-

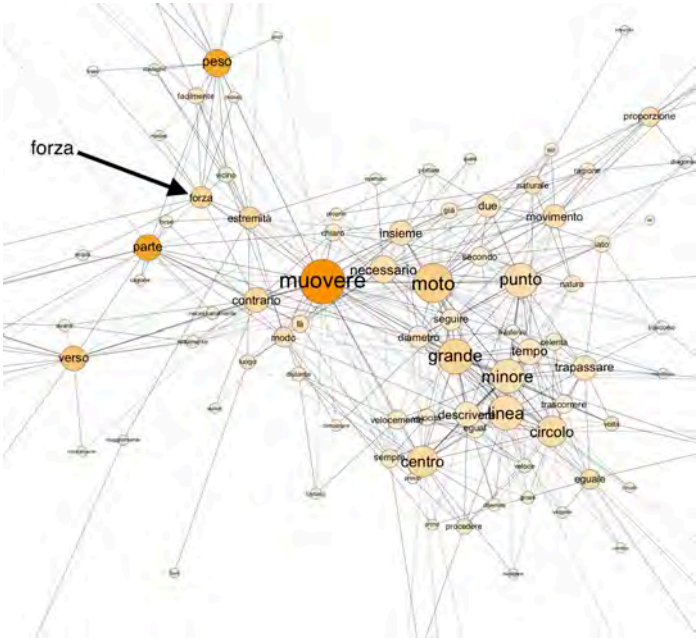


Figure 20.8: Semantic network of the 1582 Italian translation of Piccolomini's *Paraphrasis* of the *Problemata Mechanica* (central section).

> thor in these demonstrations speaks only of forces (*forze*) sustaining the weights so that they do not fall down; not about moving them.[10]

We may also note that the connection between *gravitas* and *situs* (noted above) appears in the semantic network of the term *gravis*, shown in fig. 20.10 (lower left). Here the presence of terms expressing the ideas of "ascent" and "descent" (*ascensus/ascendo*, *descensus/descendo*) also suggests Guidobaldo's engagement with the medieval theory of "positional heaviness," insofar as the texts espousing

[10]Translation Drake (in Drake and Drabkin 1969, 308). Original, full text (Pigafetta 1581, 64r): "In questa propositione si narra, che rauolgendo d'intorno à due girelle di due taglie vna corda, & quel che segue, la forza sarà vn terzo del peso, cioè se il peso sarà trecento, egli verrà sostenuto dalla possanza di cento. Direbbe alcuno ciò essere dubbioso, peroche le girelle, gli assetti suoi, le funi, & il peso della taglia di sotto fanno resistenza alla forza, & grauano sì, che ella non potrà sostenere il peso. Si risponde che queste cose ben farebbono resistenza nel mouere il peso, ma non già nel sostentarlo: & bisogna notare con diligenza che l'autore in queste dimostrationi parla sempre del sostenere solamente con le forze i pesi che non calino al basso, non del mouere."

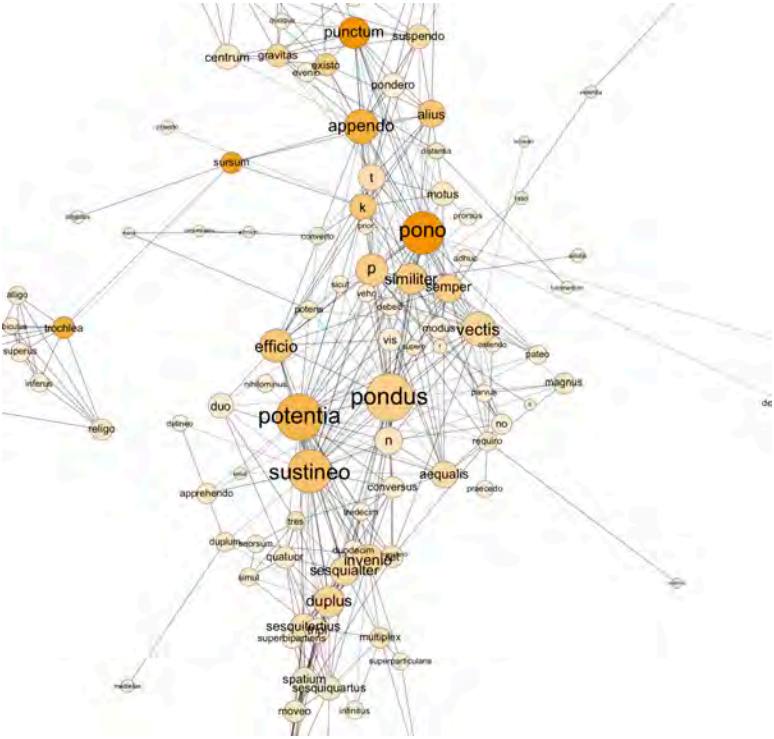


Figure 20.9: Semantic network of Guidobaldo del Monte, *Mechanicorum Liber* (central section).

this theory specified that a weight is heavier by position if its course of descent is less oblique.

I conclude this set of examples with some very brief remarks on Galileo's terminology for force. It is of course extensive and varied, and includes not only *forza*, *potenza*, and *virtù* but also other terms with a long history such as *impeto* ("impetus") and *momento* ("moment/momentum"). To illustrate the power of the semantic network approach for the analysis of his usage, we may consider two graphs generated from Galileo's *Discorsi* (1638). In fact they are components of a single graph, created by generating a semantic network for the text as a whole, then selecting the nodes for *possanza*, *potenza*, *forza*, and *virtù*. Setting a low cutoff value of 0.3, the result is two discontinuous graphs, one for the first three of these terms (fig. 20.11) and another for the last (fig. 20.12). Inspection reveals

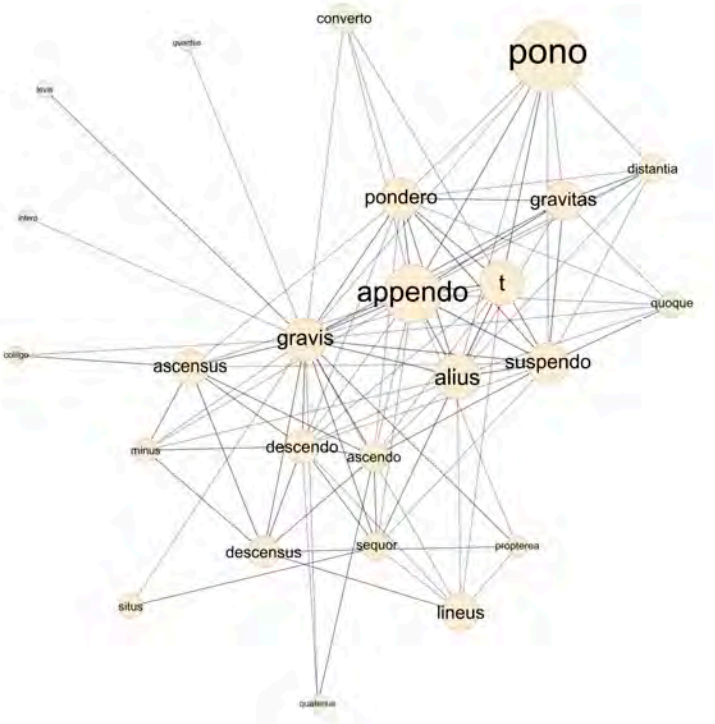


Figure 20.10: Semantic network of *gravis* ("heavy") in Guidobaldo del Monte, *Mechanicorum Liber*.

that *virtù* is much more closely associated with the cluster of terms connected with velocity and motion, while the other terms are associated with mechanical devices such as the lever and screw (*forza*) and with problems of infinite division (*potenza*). Again these results are highly suggestive and call for further investigation by close reading of the texts.

While Arboreal's current implementation of semantic analysis is provisional and needs substantial further work, these results serve as a proof of concept, showing that such an approach can capture some of the conceptual structure of scientific texts and contribute to the study of long-term intellectual developments. In each case we see that the meaning of the networks is apparent only within the framework of certain scholarly questions; it takes some knowledge of the texts to interpret these graphs, and conversely, they point the way to further topics of

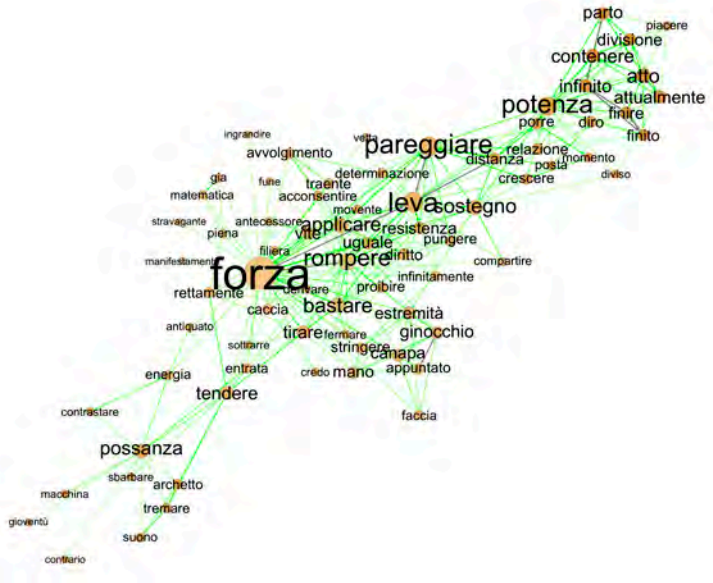


Figure 20.11: Semantic network of *potenza*, *forza*, and *possanza* in Galileo's *Discorsi*.

investigation. Considered in themselves, the semantic networks serve as a sort of "fingerprint" of the text in question, reflecting its place in the long-term development of mechanical knowledge. Moreover, this method offers the possibility of a truly multilingual approach to the history of conceptual development in science. Finally, because the method can be applied to any XML text in a language for which the necessary technology exists, it is highly general and can be used in any discipline concerned with the linguistic expression and conceptual content of textual sources. From the cuneiform archives of the third millennium BCE to the writings of Einstein, this technique has the potential to illuminate the development of human thought and to enhance our exploration of it.

Conclusions and Future Perspectives

We have much work still to do in order to complete the implementation of semantic analysis along the lines described above. One issue is that the quality of the morphological data is uneven for the various languages currently supported by our software. Insofar as the software is unable to analyze certain words in the text

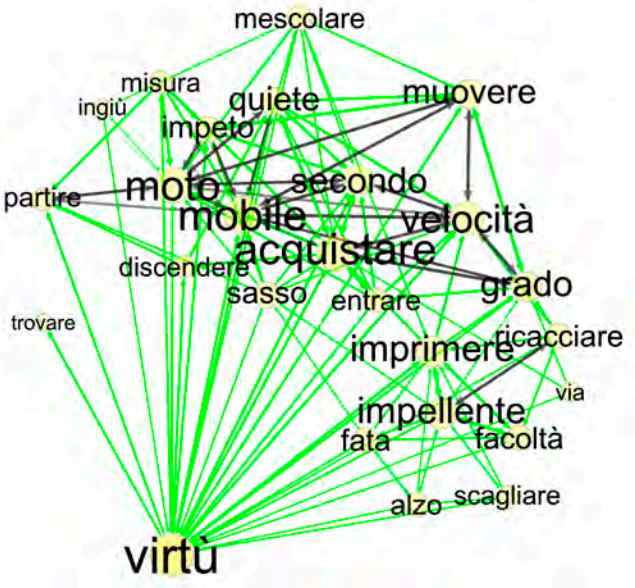


Figure 20.12: Semantic network of *virtù* in Galileo's *Discorsi*.

or refers them to multiple lemmas, the accuracy of the semantic networks that are generated is diminished. Making use of a context-sensitive part-of-speech (POS) tagger can help to avoid these problems. Yet we still have no POS tagger for ancient Greek, despite the fact that the quality of context-independent morphological analysis is very high. A related issue is that the current approach does not allow for the semantic indexing of multi-word terms such as *centrum gravitatis* or for visualizing the distribution of such terms across a text. This is so despite the fact that Arboreal provides powerful functionality for extracting and tagging such terms. We will remedy this deficiency in the near future. It would also be highly desirable to be able to navigate back into the text by clicking on nodes in a graph, which should be a straightforward function to implement.

Two more fundamental challenges remain if the full potential of Arboreal is to be realized. First, the maintenance and stability of the code base requires constant attention. Although the choice to implement Arboreal as a Java application enabled us to avoid many server-side maintenance issues, the current version consists of approximately 30,000 lines of code, and it is a significant challenge to ensure that it conforms to the latest version of Java and works on all common plat-

forms. For the foreseeable future the latest version of the code will be available on the Harvard Archimedes server (http://archimedes.fas.harvard.edu/arboreal).

There is no doubt that the complexity of Arboreal has been an impediment to its adoption by the scholarly community. Indeed, the most successful examples of its use have been in contexts where one or more of the developers themselves were available for consultation. While the history of Arboreal's development demonstrates the importance of close collaboration, it also points to the need for better documentation and communication. A software package needs to be self-sustaining if it is to be broadly adopted. There is therefore an urgent need to simplify the software where possible and to provide better documentation of different usage scenarios. We will prioritize these tasks in the near future.

In closing, I believe that the history of Arboreal gives good reason to be optimistic about its future. Arboreal has survived many different versions of Java and many operating system updates and is still a going concern, due to the dedicated efforts of many people in various institutions. I believe that a major reason for its continued vitality lies in the way in which it embodies Peter Damerow's compelling vision of the role of information technology in the humanities. In this vision the power of computational techniques is harnessed as far as possible, but they are not treated as ends in themselves; the goal of software design is to enable researchers to engage interactively with technology and with the sources that they study; and the goals of simplicity and open access are of supreme importance. Although the challenges that remain are as as great if not greater than ever, there is every reason to believe that this vision will continue to be relevant for the foreseeable future, and that with further work we will come even closer to realizing it.

Acknowledgments

I would like to express my deep gratitude to all who have been involved in the development of Arboreal over the years; to Jürgen Renn for inviting me to Berlin in 1998 and for his steadfast support ever since; and to Matthias Schemmel for his friendship and patience as I worked to complete this paper. I am honored to be able to offer it as a small tribute to Peter's memory.

References

Anderson, Chris (2008). The End of Theory: The Data Deluge Makes the Scientific Method Obsolete. *Wired*. URL: http://archive.wired.com/science/discoveries/magazine/16-07/pb_theory.

Damerow, Peter and Jürgen Renn (2010). *Guidobaldo del Monte's Mechanicorum Liber*. Berlin: Edition Open Access. URL: http://edition-open-access.de/sources/1/index.html.

———— (2012). *The Equilibrium Controversy: Guidobaldo del Monte's Critical Notes on the Mechanics of Jordanus and Benedetti and their Historical and Conceptual Backgrounds*. Berlin: Edition Open Access. URL: http://edition-open-access.de/sources/2/index.html.

Drake, Stillman and Israel Edward Drabkin (1969). *Mechanics in Sixteenth-Century Italy: Selections from Tartaglia, Benedetti, Guido Ubaldo, & Galileo*. Madison: University of Wisconsin Press.

Drake, Stillman and Paul Lawrence Rose (1971). The Pseudo-Aristotelian *Questions of Mechanics* in Renaissance Culture. *Studies in the Renaissance* XVIII:65–104.

Hyman, Malcolm (2007). Semantic Networks: A Tool for Investigating Conceptual Change and Knowledge Transfer in the History of Science. In: *Übersetzung und Transformation*. Ed. by Hartmut Böhme, Christof Rapp, and Wolfgang Rösler. Berlin, New York: De Gruyter, 355–367. Originally published in 2006 as Preprint 320 of the Max Planck Institute for the History of Science.

Hyman, Malcolm and Jürgen Renn (2012). Toward an Epistemic Web. In: *The Globalization of Knowledge in History*. Ed. by Jürgen Renn. Berlin: Edition Open Access. URL: http://www.edition-open-access.de/studies/1/36/index.html.

Pigafetta, Filippo (1581). *Le mechaniche dell' illustriss. sig. Guido Ubaldo de' Marchesi del Monte: tradotte in volgare dal sig. Filippo Pigafetta*. Venice: Sanese.

Schiefsky, Mark J. (2007). New Technologies for the Study of Euclid's *Elements*. URL: http://archimedes.fas.harvard.edu/euclid/euclid_paper.pdf.