CHAPTER 8

Using Quantitative Methods for Measuring Inter-Textual Relations in Cuneiform

M. Willis Monroe

Introduction

One of the many benefits of quantitative methods in the digital humanities is their ability to test intuitions and assumptions about a body of textual material quickly and efficiently. This is often most useful when the corpus is extremely large, and regular methods of reading or analysis are not possible. While quantitative methods can provide a level of certainty to results or conclusions, their use in exploratory hermeneutics is equally valuable.¹ This chapter intends to explore the viability of these methods in order to investigate a small corpus of Babylonian astrological material concerned with the association between ingredients used in medical treatment and the signs of the zodiac. The methods employed are relatively simple and can be performed using free open-source software, yet they allow for an iterative process of knowledge creation and investigation that produces multiple results at different stages of the research.² In the spirit of this volume, a significant amount of explanation will be included to break down the techniques and methodology behind the processing and analysis of the texts.

The underlying goal of this analysis is to understand the scholarly practices behind the composition and editing of some unique examples of Late Babylonian astrology. The period under Achaemenid and Hellenistic rule in Babylonia (late sixth to second centuries BCE) saw a flourishing of scholarly activity recorded on cuneiform tablets in southern Iraq. Astronomy and astrology were, in particular, realms experimentation and rapid development. The texts under consideration in this chapter comprise one of many new types of astrology, each borrowing and building on previous forms of knowledge. The

---

¹ “There is a tendency in debates to reduce the potential of computation to a methodology of quantification. The nature of digital humanities is hybrid, however, and there is not an a priori discontinuity with the hermeneutic traditions” (Zundert 2015, 340).

² For iterative processes (developed in C++) and for projects on object typologies, especially to set up qualitative-data processing, see in this volume, Martino and Martino, 120.
material represents a collection of tables of astrological significance; the area under investigation is a row of medical ingredients linked with pairs of zodiacal signs. The text is called the “Micro-zodiac,” so named for its use of a smaller set of zodiacal signs to subdivide the regular 12 signs of the traditional zodiac.3 The traditional Babylonian zodiac is the antecedent to the western zodiacal signs, as illustrated in the following table:

| Babylonian sign | Translation  | Modern equivalent |
|-----------------|--------------|-------------------|
| ḪUN             | Hired man    | Aries             |
| GU₄.AN.NA        | Bull of heaven | Taurus            |
| MAŠ.MAŠ         | Twins        | Gemini            |
| ALLA            | Crab         | Cancer            |
| A               | Lion         | Leo               |
| ABSIN           | Furrow       | Virgo             |
| RIN₂            | Scales       | Libra             |
| GIR₂.TAB        | Scorpion     | Scorpio           |
| PA              | Pabilsag     | Sagittarius       |
| MAŠ₂            | Goat-fish    | Capricorn         |
| GU              | Great one    | Aquarius          |
| ZIB.ME          | Tails        | Pisces            |

The methods employed here are often used with much larger and more complete corpora that represent bodies of material beyond the ability of a single scholar to read closely. In those cases, the methods allow for a process of analysis not otherwise feasible. Widely used in the field of Early Modern Literature, computer-aided textual analysis can answer questions about themes or the usage of terms across a corpus containing millions of words. The corpus with which this paper is concerned differs in that it is small enough to read closely, but it is lacking in coherence. The individual words within the corpus are grouped in clusters on the text and lack any linguistic syntax. The damaged and incomplete nature of the corpus makes traditional forms of reading and analysis difficult. So a quantitative method allows for more concrete conclu-

---

3 The Micro-zodiac series was the subject of my dissertation research (Monroe 2016).
sions or insight into the organization of the text obscured by its format. Through computer-aided textual analysis, we can visualize what would otherwise be mere hunches gleaned from a close reading of the texts.

Modern techniques in textual analysis are often used to perform a so-called “close reading” of a textual corpus, a method of analyzing the constituent parts and makeup of a text to reach a deeper understanding. By leveraging the analytical power of digital methods, the researcher can survey an entire corpus of material for a particular term or area of interest, or, in other cases, allow the material to speak for itself. This chapter will attempt to show another use for these methods. The corpus used here is smaller than most of those used in other forms of textual analysis, but the disjointed nature of its content is immediately analyzable through similar methods. The traditional methods of textual analysis often remove words or terms from their local context and analyze them as a whole, looking for patterns among groups of words represented in documents or manuscripts. In the case of the Micro-zodiac, the first step is already completed; the lack of any syntactic grouping allows for the bunches of terms within each cell of the tables to represent a “document” in our analysis.

Theory

Much of the theory behind the methods employed in this chapter in particular and in the field of digital humanities in general were first concretely laid out by Franco Moretti in his book *Graph, Maps, Trees*, in which he advocated for a form of “distant reading.” The “distant” in “distant reading” need not refer to a lack of physical proximity between the reader and text; it refers, rather, to methods of analysis that involve mapping and the distance between objects of the text (or texts). The reader or scholar is still very close to the text, in some cases using minute elements of the text as the unit of their analysis. The overall

---

4 The corpus of the Micro-zodiac texts consists of 15 known tablets of varying degrees of preservation. The most complete have more than 20 cells (the unit of analysis for this paper) of ingredients on them. The dataset contains 476 individual ingredients pulled from the preserved cells on the known Micro-zodiac tablets.

5 Jänicke et al. 2016, 2.

6 These techniques are increasing rapidly in frequency of use within the digital humanities. A recent survey of journal articles charted a large increase in papers from 2005 to 2014 that used techniques classified as “distant reading” (Jänicke et al. 2015, fig. 3).

7 Moretti 2007.
aim of the methodology employed here is to produce a graph used to visualize the connections between cells. A graph is made of nodes and edges, both of which are determined by the nature of the evidence. In many cases nodes represent texts within a corpus or documents within a genre. Edges represent connections between nodes, often formed by shared terms, words, or other meaningful units of analysis. The choice of what represents a node or an edge can change as the analysis proceeds; in the method that follows nodes and edges are processed one way and then inverted for another view of the same data.

The methods of distant reading are most commonly divided into two categories, “supervised” and “unsupervised.” The distinction between the two is at times fluid, and there are many overlapping terms that encompass one or both, but at its simplest level, the difference between “supervised” and “unsupervised” methods of text analysis involves whether or not the researcher has chosen to “classify” the material. The processing of classifying textual data involves identifying important features, for example, length, word counts, and format, that can be represented as numbers and ascribing these features to certain classes of documents. Analysis with a text that has already been partially classified is considered to be supervised, whereas an analysis with an unclassified dataset is unsupervised. For instance, a study might involve attempting to associate similar authors with an unattributed new trove of anonymous novels in order to discover the unnamed author. A supervised method of analysis would involve classifying sets of novels (by their features) for which the authors are known, comparing the new material against the known corpus, and then assigning it to one of the existing groups. This is a supervised method because the researcher is supervising, commonly called “training,” the algorithm by providing it with known data prior to performing analysis on the unknown material. The unsupervised method would involve not classifying any of the existing corpus, but, rather, letting the algorithm decide which records (i.e., novels) out of the entire corpus belonged together. Here, defining features is also important, but it is crucial to note that the documents are never grouped into classes based on these features; that task is left to the algorithm with the hope

---

8 “Today, data mining often implies unsupervised learning, whereas machine learning is more commonly applied to supervised learning processes. But this boundary can be drawn in several different ways: sometimes data mining names the practice that corresponds to machine learning’s theory” (Jockers and Underwood 2015, 292).

9 Supervised method: more generally, a method in which the researcher has chosen to “classify” part of the material before the analysis.
that the known works of certain authors would be found in similar groups. The benefit of the unsupervised method here, however, is that it breaks what are otherwise researcher-oriented biases and presents an “objective” view of the material, free from prior classification. Perhaps the unsupervised method detected that a group of novels thought to belong to one author most closely match the unknown material in our example. The conclusion is up to the researcher here, but the unsupervised method can offer insight into a dataset without any preconceived notions of structure or definition. Unsupervised methods are commonly used to explore a dataset, after which a supervised method can be used to hone in on a particular area of interest.

Text Background

The Micro-zodiac material comes from Hellenistic Babylonia, where it was written by scribes educated and working in the Late Babylonian scholarly communities of southern Mesopotamia. In this period local rule in Mesopotamia had been eclipsed, and the cities of Babylonia were part of the wider Seleucid Empire. The scribes were attached to the temples through cultic roles and family ties. These scholars wrote extensively on many forms of traditional Mesopotamian scholarly knowledge, from divination to ritual and from mythology to astrology. They have taken their place in our modern history of science as skilled mathematical astronomers; their methods for calculating the planetary ephemerides and solar and lunar eclipses were renowned throughout the ancient world. Terms and constants computed in Babylonia show up in other traditions, and the ethnonym “Chaldean,” referring to Babylonian scholars, became synonymous with ancient astronomers/astrologers of the Middle East.

---

10 An important caveat here is that the output generated by unsupervised methods is only as good and as accurate as the input data. An unsupervised method could very well suggest theories and connections between points in the data that are not borne out by the evidence because particularly crucial differentiating variables were left out of the analysis. See in this volume, two other examples of unsupervised methods: Artificial Neural Network (ANN) algorithms applied to landscape archaeology by Ramazzotti (63–65), and the multivariate analysis of figurines by Martino and Martino (118). As for a supervised example (on cylinder seal imagery), see also in this volume, Ludovico (92–95).

11 Jockers and Underwood 2015, 294–295.

12 Useful introductions to Babylonian astronomy and astrology can be found in Rochberg (2004) and Steele (2008).
Along with performing mathematical calculations, these same scribes wrote a large corpus of astrology that exhibits a similar complexity and development of thought. These texts take the configuration of planets and stars and compute new meaning and significance based on their positions and movement. Much of their work borrowed from a long tradition of astrology going back roughly two millennia to the earliest forms of Mesopotamian divination. A large astrological compendium of omens, called *Enūma Anu Enlil*, formed the basis of much of their education. During this period, however, these ideas were modified and further developed, producing new types of texts that experimented with novel methods of association to produce new meaning. The Micro-zodiac series in particular shows how threads of previous forms of knowledge were compiled and woven together using the recently invented zodiac as their organizing principle.

The Micro-zodiac tablets come from the cities of Babylon and Uruk and date roughly to the third and second centuries BCE. Each tablet preserves one or more (most have two) tables of astrological knowledge. It is important to note that these are formal tables demarcated by incised lines separating the columns and rows from each other. Often the tables have a header containing information about celestial omens related to the lunar eclipse. The examples from Uruk are known for their fabulous imagery of the zodiacal signs located in a horizontal band above the table itself. The material is organized around the zodiac—hence the previously mentioned imager—and each table is associated with one of the 12 signs of the zodiac. The twist in the Micro-zodiac is the subdivision of each sign into twelve more signs, each of which governs only 2.5° of the entire sky or year. Each of the subdivisions represents one of the twelve columns within each table. In a sense the entire series is a large spreadsheet, with each table representing one-twelfth of the overall content. Tablets from the series generally have two tables, and therefore one-sixth of the entire text. The two tables are always in sequence, starting with an odd-numbered sign on the obverse of the tablet and the subsequent even-numbered sign on the reverse. Returning to the organization, each column of the entire text can

---

13 The known exemplars of the series are currently housed in the Iraq Museum in Baghdad, The British Museum in London, the Louvre in Paris, and the Vorderasiatisches Museum in Berlin. Only one exemplar is derives from controlled excavation (by the Germans in Uruk). The rest are from uncertain provenance collected during the late nineteenth and early twentieth century by the aforementioned museums. It is reasonably certain that the known exemplars all come originally from either Babylon or Uruk.

14 1° of the Sun’s motion through the zodiac is roughly equivalent to 1 day of the ideal year. The sun travels 360° around the earth in a year, and there are 360 days in an ideal Babylonian calendrical year.
be identified by the zodiacal sign to which its table is assigned, hereafter the “Major sign,” and the zodiacal sign to which the individual column has been assigned, hereafter the “minor sign.” In simpler terms, each column shares a Major sign with the other 11 columns in its table and has a minor sign that is unique within its table. This system of Major-minor sign pairs creates a unique pattern of 144 couples, each of which defines the location of a column within the entire text.

The Micro-zodiac table itself contains four major rows of content: medical ingredients, celestial divination, a cultic calendar, and daily advice. Each row has many parallels in existing textual traditions; the final row, concerning daily advice, for example, borrows much of its content from the well-known hemerological texts, including the Babylonian Almanac. For this paper, however, only the first row, the medical ingredients, will factor into the analysis. The rows interact with the organization of the columns in different ways. The middle two rows, divination and cult, repeat their content every twelve columns. This means that, in a sense, their contents only depend on the latter of the two zodiacal signs, the minor sign, which indicates the column. This pattern restricts the total amount of content in these two rows to only 12 possibilities.

The first and final row, the medical material and daily advice, respectively, do not show any discernable pattern. Under the same exact column on another text, the two cells will contain the same material, i.e., material found under one pair of signs will be the same as material found under the same pair on another tablet. However, the material does not repeat on each table under the same minor sign, as it does with the middle two rows. The fact that material is found on a different table under the same unique pairing of major and minor signs suggests that the inclusion of material in this row does follow a predictable pattern within the context of the Micro-zodiac.

Closely related to the Micro-zodiac texts are another genre of astrological material called Calendar Texts; these tablets contain much the same information organized under the same system as the Micro-zodiac but are structured in a slightly different way. They factor into the analysis below because the content of the medical-ingredients section follows the same rules and parallels the Micro-zodiac material.

The medical ingredient cells contain four to six individual items. The first is always a location, either a city or a temple. The next three ingredients are always found in the following order: wood, plant, stone. However, while some texts stick to a rigid set of four total ingredients, others might have double the number of ingredients in one or more of these three categories. Often there are

---

15 A good overview of Babylonian medicine can be found in Geller 2010.
two ingredients listed for the final category, stone. Below is an excerpt from the middle of the Leo side of tablet VAT 7847 (text 7 in the corpus). Here, because of the location of Leo in the zodiac (the fifth sign), the boundary between Pisces and Aries (the last and first signs, respectively) is found in the physical middle of the tablet itself. The two cells list cultic buildings first, the Urinnu and the Ekur in Nippur. Next they list a sequence of ingredients following the paradigm of place, wood, plant, and stone.

Table 8.2 List of ingredients

| Leo-Pisces   | Leo-Aries       |
|--------------|-----------------|
| E₂ U₄.RI₂.IN | NIBRU⁹ᵏ¹ E₂.KUR|
| gišKUR.RA    | gišES₂₂         |
| u₂sī-ḫa      | u₂A.ZAL.LA₂      |
| na₄BAL       | na₄AN.ZAH       |

Most of the ingredients found within these rows are well-known from the existing medical tradition in Mesopotamia. The woods, plants, and stones were used to create amulets and phylacteries employed in the treatment of illnesses. It is most likely that the first ingredient in each column, listing religiously significant places, refers to the dust from that location used in treatment. Prior to the Late Babylonian period, these ingredients could be found in the therapeutic corpus used in the construction of various remedies for ailments. In fact, the categorical scheme is well known from medical texts contemporary

---

16 VAT 7847 was first edited by Ernst F. Weidner in his 1967 edition of the first identified Micro-zodiac texts.

17 The bit urinnu (E₂ U₄.RI₂.IN) was a generic term for a temple treasury (CAD, vol. U, 227, urinnu B). gišKUR.RA is found in the lexical lists and must refer to a “mountain-tree,” as resin of the sīhu-tree was used in recipes (CAD, vol. S, 242, sīhu d). The E₂.KUR in Nippur is the chief temple of the god Enlil, the šiqdu-tree (ES₂₂) is probably an almond (CAD, vol. Š/III, 94, šiqdu), the azallû-plant is well known from the medical corpus (CAD, vol. A/II, 524, azallû), and finally anzahhu-stone is a glass-like material also used in medical contexts (CAD, vol. A/II, 154, anzahhu). Very few of the ingredients are translated in this chapter, partly because their identification is not certain, but also because their actual correspondence with any botanical name is not important for the method employed here.

18 Note the parallel with the text edited by Nils Heeßel (2005), in which a similar group of four medical ingredients is included, with the last one in each case being the “dust” (saḫar) of a location.
with the Micro-zodiac. Interestingly other texts using this scheme order the categories in slightly different ways: the texts that Heeßel has published use the order: stone, wood, plant, place. This scheme seems to be well known; in fact, one text in particular directly references these groupings of ingredients as a medical device, referring to the generic terms for the ingredients “stone, plant, tree” in connection with the zodiacal schemes. It should be noted that the inclusion of multiple related zodiacal schemes in this tablet suggests that the associations present in this text between medical ingredients and signs of the zodiac constituted a paradigmatic structure that existed outside of the Micro-zodiac series.

The guiding question behind this study was whether there was an underlying system of organization centered around these medical ingredients and their association with certain zodiacal signs. Perhaps certain signs were linked in ways that were not immediately obvious; for instance, a medical ingredient might be linked to a certain minor sign and only appear when that sign was present within the column’s major-minor sign pair. If the other rows displayed some form of organizational structure, albeit quite simple, perhaps the medical-ingredients row could also contain an underlying logic—one not immediately visible. A good example of this practice is an astrological text written by the scribe Iqīša, which links signs of the zodiac with medical ingredients made from the animals associated with the zodiacal signs. In particular, the initial assumption of this method is that the organizational structure will be based on one of the two zodiacal signs that identify the location of each column, similar to two of the other rows. Astrology of this period is full of patterns and repetitive sequences, many of which found their way into Hellenistic geometric descriptions of astrological significance. Some of these patterns date to earlier periods; for example, the linking of diseases and signs of the zodiac was shown by Mark Geller to be a development based on calendrical medicine from the earlier Neo-Assyrian period.

The theory of borrowed associations and existing paradigms can easily be tested with other shorter contemporary texts. When the ingredients and

19 Heeßel 2008, 9.
20 This text, LBAT 1593, refers to both the Micro-zodiac and Calendar text schemes as “the animals of 13 and 4,37” (Reiner 2000, 424).
21 Steele 2011, 338–339.
22 Rochberg-Halton 1988.
23 “The profound change taking place is that a traditional hemerology-based system of favourable and unfavourable days of the month for various rituals (STT 300) has been replaced by zodiac-based system which assumes astral influences over the same spells and rituals (BRM 4 20 and 19)” (Geller 2014, 57).
schemes within the Micro-zodiac are compared to other known texts such as the aforementioned material edited by Heeßel, there is a lack of parallel content; this suggests a familiarity with material but not the adoption of an organizational scheme. For instance, many of the medical ingredients are similar, and the methods of association, i.e., the idea of linking ingredients to zodiacal signs, are clearly related, but the connections themselves are not shared between corpora. Another astrological text links signs of the zodiac with cities, many of which are found within the Micro-zodiac, but under different signs. Making this type of comparison is relatively easy to do, as the external material to the Micro-zodiac is more concise and simpler in format. The question remains whether or not there are other forms of organization underlying the material in the Micro-zodiac that are not detectable by comparison with other existing textual exemplars. If there were, we would assume these connections would not be bound to the Micro-zodiac scheme, but rather would present themselves through repetitive usage of ingredients in different sign pairings.

The methodology outlined below allows the text to be analyzed in a way that should bring to light any schemes of organization that are not immediately obvious. By analyzing the network of ingredients as a whole, we can highlight areas of commonality between cells. Crucially this would suggest the existence of a scheme without needing to have a prior example of the text returning to the difference between supervised and unsupervised methods, this method of generating groupings and structures *ex novo* without needing *comparanda* is one of the benefits of the unsupervised technique. It is important to note, however, that it is not without its pitfalls. The methods employed below are entirely dependent on the quality and form of the data entered. The unsupervised methods in particular may produce results that do not reflect a true understanding of the material, without the types of regulatory metrics that derive from a more thorough understanding of research questions, as is the case with supervised methods.

**Methodology**

Since one of the aims of this volume is to explain some of the methods of the digital humanities in understandable terms, it is worth running through the process of going from text to analysis in detail. The following section will run

---

24 Steele 2015.

25 The following methods were all completed with a digital “notebook” using the IPython platform. This allows the researcher to work iteratively through processing their data,
through the steps in turn, explaining the methods and terminology behind each step. Parallel to the explanation, I will include a small subset of the data to serve as an illustration of the steps. We will start with the text objects themselves and then progress through philological work, digital coding, and finally analysis and visualization. This process spans both work in museum collections as well as digital processing with the Python programming language.

It is important to state at the outset that none of this work would be possible without very traditional philological study. In order to begin the work, the data had to be collected. The texts were located in museum collections, visited, photographed, studied, translated, and finally edited. The output of this work, of course, has many uses, and it forms the bedrock of any future study of the material, this investigation being only one part.

After the texts have been edited, the relevant sections have to be selected and converted into a format that allows for easy manipulation by data analysis tools. In this case I transcribed the data into a spreadsheet. It is important to choose a format and method of coding that is both flexible and standardized from the outset. This allows for any type of textual data to be represented in a form that is translatable and readable by later stages of your analysis. Each row represented one ingredient, and the columns were laid out as follows (broken into two lines here):

There are more columns than necessary here, but it is best to strive for completeness rather than to have to go back through the dataset to re-enter a crucial bit of information later on. The redundancy, in particular between the signs and numbers, also makes things like labels for visualization easier later. A few notes on the columns: “Index” serves as an overall index of every row; “text” represents an internal count of the texts in the corpus; “text_type” records the type of text, allowing for future expansion; “face” refers to the physical side of the tablet; “major_sign,” “major_number,” “minor_sign,” “minor_number,” and “zodiacal_location” locate the ingredient on the text; “raw” represents the ingredient as it appears in the transliteration (the Latin-character equivalent of the cuneiform script); “name_only” removes the determinative; “type” is

26 One important facet to the development of techniques in digital humanities is their ability not to replace but to “enhance” the traditional work of text scholars (Jänicke et al. 2015).

27 A good description of digital encoding of textual data can be found in Sinclair and Rockwell’s chapter, “Text Analysis and Visualization” (Sinclair and Rockwell 2015, 279).
essentially a translation of the determinative; “translation” is an English translation when relevant; “damage” and “unclear” are just flags for when the raw transliterated string is damaged or an unclear reading—these can be used later on to restrict the dataset to certain confidence levels—and notes is just for internal comments about the data.

Once the data is in a digital form, it can then be imported into any number of tools; in this case the programming language Python was used, as well as the Pandas and NumPy libraries for data processing.\textsuperscript{28} The Pandas library\textsuperscript{29} functions around DataFrames, in essence a form of digital table. They come with their own built-in methods and tools for managing and viewing the data. Pandas allows for logical operations to be applied quickly and efficiently to the entire dataset. First, any cells with no relevant data in the “name_only” field were removed; these included cells where there was no preserved information beyond a determinative, or where there were unclear readings of ingredients. Next a few custom columns were constructed out of previous columns: a unique location identifier that combined the text number and the two zodiacal locations, and a composite column including the name of the ingredient and its type. At this point, because I included the “text_type” field I was able to easily combine or separate out various groups of texts into new DataFrames for later analysis. For instance, I could analyze the Micro-zodiac material by itself or include the Calendar Texts as well.

\begin{table}[h]
\centering
\begin{tabular}{llllllll}
\hline
Index & Text & text_type & face & major_sign & major_number & minor_sign & minor_number \\
\hline
1 & 1 & Micro-zodiac & Obv & Aries & 1 & Cancer & 4 \\
\hline
zodiacal_location & Raw & name_only & type & translation & damage & unclear & Notes \\
\hline
1-4 & bar₂- & bar₂-sip₂ & place & Borsippa & & & \\
\hline
\end{tabular}
\caption{Example row from the input data}
\end{table}

\textsuperscript{28} McKinney 2010, \url{http://conference.scipy.org/proceedings/scipy2010/} (accessed July 2, 2017); Walt, Colbert, and Varoquaux 2011.

\textsuperscript{29} A “library” is a collection of commonly used programming functions that are distributed as a package for use by a wide range of users.
**Figure 8.1** Counts of ingredients in the Micro-zodiac
As mentioned above, the process of producing and analyzing data is an iterative process that generates results while moving toward an end research goal. At this point the coded and imported data can already be processed and visualized in interesting ways. The power of using scripting languages and their associated libraries is that it is easy to explore the data in order to produce a general sense of its shape and make-up. Using some simple logic, I was able to produce a graph of the counts of various ingredients within the entire corpus using the Matplotlib and Seaborn libraries.30 In this case I selected ingredients that appeared more than four times in the entire corpus and graphed them against each other.

This type of counting and visualization is feasible by hand, but the power inherent in this technique is that while one prepares the data for further, more complicated analysis, these intermediate steps can help one explore the data and refine the research question while developing further methods. This initial count gave me a sense of the most common ingredients in the corpus. It is interesting to note the low total numbers for all of these ingredients, yet the numbers are high enough to suggest that there could be shared content between different cells. If there were no ingredients with higher counts—if all ingredients had a count of one, for example—then it would be clear that there would be no value in graphing the connections between cells. The inverse is also true: if the counts were too high, it would suggest that the total pool of ingredients was too small and too interconnected to make a meaningful differentiation between groups of interconnected cells. Also notice that the third ingredient from the right is $e_2 \times$, meaning an unknown temple or cultic building. This is an obvious example of where damage or unclear readings can introduce errors into the processing. All of the instances of this ingredient are considered the same and if not accounted for may lead to false similarities between cells later on in the analysis.

After the data had been cleaned and new columns added, it was run through a number of functions to perform the analysis. The first step was to form bags of words for each cell in each text. The bag-of-words model is a well-known method in textual analysis in which all the individual lexemes from a text are flattened and placed in a large group.31 Syntax and order are lost and do not figure into the final analysis, but this model offers a level of simplicity that functions very well for comparing the similarity of texts within a corpus.

---

30 Hunter 2007; Waskom et al. 2016, <https://zenodo.org/record/54844> (accessed July 14, 2017).
31 See also in this volume, Svärd, Jauhiainen, Sahala, and Lindén (247–248), who applied the Continuous Bag-of-Words model to semantic analyses.
The table below is an example of the bag-of-words model for three cells from the dataset. Each cell is given its own row, and the ingredients are listed in one line. While this model looks relatively simple for this dataset, most applications of the bag-of-words model involve listing the entire contents of, for instance, a novel, in one line.

**Table 8.4  Bag-of-words model for three cells**

| Cell   | Items                        |
|--------|------------------------------|
| #3/4-11a | e₂ nin.gir₂.su, dup-ra-nu, ak-tam, mar-ḥal-lu₄ |
| #6/4-11  | dup-ra-nu, ak-tam, mar-ḥal-lu₄          |
| #7/5-8   | sig₄ unug, dup-ra-nu, ši-im-ra-nu, nir₂   |

a  The convention used throughout this chapter for naming cells is as follows: the character “#” followed by the number of the text within the corpus, then a forward slash, and finally a pair of numbers representing the Major and minor signs of the zodiac, respectively. “#3/4-11” stands for the third text in the corpus and the cell assigned to Cancer-Aquarius.

With a bag of words (in this case ingredients) for each cell, the entire group could then be run through a function to produce a Document Term Matrix (DTM). The matrix is a large “vector space” in which the cells reside. Each unique term within the corpus becomes a dimension on which the documents are plotted; for instance, if your corpus consists of documents that only use three words, the resulting vector space would be recognizable as a three-dimensional plot. In practical terms the DTM is represented by a large table in which each row is one of the input documents, in this case a cell from one of our texts, and each column represents a unique term that shows up somewhere in the entire corpus. The function, CountVectorizer from the Scikit-learn module, runs through each bag and tallies a total list of all terms, then

---

32 See also in this volume, Eraslan (284–285, 305), on vector space and interoperability.

33 See also in this volume, two other examples of structured matrices used for identifying relationships: Auto-Contractive Map (Auto-CM) applied to landscape archaeology by Buscema in Ramazzoti (68 fig. 2.1a, 73n36, 74), and textual correspondence analysis for the investigation of cylinder seals by Ludovico (94–97).

34 CountVectorizer was chosen over another popular method in textual analysis called Term Frequency–Inverse Document Frequency (TF-IDF). The latter is quite popular because it evaluates the importance of terms within one text in inverse relationship to their frequency in the entire corpus, essentially identifying the terms that are most important for identifying a document. This method is based on frequency counts for terms within
marks, for each bag, which terms appear in that bag. The end result is a matrix of all cells and all terms, with the count for each term in each cell tallied in the table. The table below is a DTM for the example given above. As you can see, the columns consist of a list of every term that exists in the example corpus. Each row is one of the cells in our example corpus. The values of the cells in this table are the number of occurrences of a term in a cell, in this case either zero or one.

| ak-tam | dup-ra-nu | e₂ nin.gir₂.su | mar-ḫal-šu₄ | nir₂ | sig₄ | unug | ši-im-ra-nu |
|--------|-----------|----------------|-------------|------|------|------|------------|
| #3/4-11 | 1 | 1 | 1 | 1 | 0 | 0 | 0 |
| #6/4-11 | 1 | 1 | 0 | 1 | 0 | 0 | 0 |
| #7/5-8 | 0 | 1 | 0 | 0 | 1 | 1 | 1 |

Immediately some patterns begin to emerge; it is clear that the term *duprānu* is present in all cells, and the last three terms are only present in the final cell. This example is relatively simple and compact, but on a real corpus this step produces a huge dataset; the resultant table must include a cell for every word in the corpus and every document. On a small, example dataset such as the one above, the result is readable and understandable, but a real DTM is of little use without further processing.

With the matrix of terms prepared for the entire corpus, a function can be applied to the data to find the similarity between cells based on their shared terms. In this case the metric used was one called “cosine similarity.” This is a geometric method for determining similarity between rows in the DTM by assigning each row to a vector in multi-dimensional space and calculating the cosine of the angle between each vector. This method is well suited to the sparse binary nature of this dataset. Because the values from the DTM are either one or zero, representing the presence or absence of a term, respectively, the cosine of the angle is best suited for an analysis of similarity. If, on the other hand, we were concerned with the number of terms within each cell and documents and within the corpus as a whole. It works well when there are many terms shared among all documents, as it weeds out common terms and focuses on the important rare terms. Because all terms in the present corpus are important, and counts greater than one are rare, a simple count of terms was used.

Pedregosa et al. 2011, <http://jmlr.org/papers/v12/> (accessed July 5, 2017).
took the frequency of terms to be an integral part of the difference or similarity between cells, then another method (for instance, Euclidean distance) would be better suited. With the cosine similarity computed across the entire corpus, a matrix can be constructed that shows each cell and its similarity with every other cell within a range of zero to one (with one being identical). Like the graph above of counts of ingredients, this step can be graphed as well, giving us a quick window into the data.

From the graph above (Fig. 8.2), you can see that only a few of the cells, on the far right, are exactly identical to each other. The majority of cells that have any similarity at all (this graph does not include cells with zero similarity) are relatively dissimilar. Applied across the entire corpus, this method produces an adjacency matrix, essentially a table in which the rows and columns are cells from the corpus, and at the intersection of each row and cell, a number is given for how similar the two cells are. As expected there is a diagonal line of 1’s running down through the middle of the table where each cell is exactly similar with itself. When the similarity is computed across the entire example dataset, the following table is the result:

**TABLE 8.6 Adjacency matrix for example dataset**

|     | #3/4-11 | #6/4-11 | #7/5-8 |
|-----|---------|---------|--------|
| #3/4-11 | 1       | .866025 | .25    |
| #6/4-11 | .866025 | 1       | .288675|
| #7/5-8  | .25     | .288675 | 1      |
This table shows us that each cell is similar to itself, which is expected. The more interesting results are at the intersections of different cells, where we find differing counts. From this example dataset, it is clear that the cells located under Cancer-Pisces in texts 3 and 6 are very closely related, sharing three terms between them, resulting in a cosine of .866025. Meanwhile, the cell under Leo-Scorpio from text 7 shares only one term with the previous two cells, which, if you return to our DTM table above, was the term duprānu. This results in two different cosine values because texts 3 and 6 differ in length. Text 7 is more similar to text 6 than is text 3, because text 3 has one more term than does text 6, and text 7 shares only one term with both. As with the production of a DTM, this example dataset is easy to understand; on a real dataset the result would be too large to comprehend. This is where the visualization of data is necessary for understanding the results. The adjacency matrix produced above can then be exported to a Comma Separated Value (csv) file and imported into a graphing tool such as Gephi.36 The procedure for creating data that is digestible by graphing programs is somewhat opaque and is left as an exercise to the reader. Gephi in particular will silently reject data in many cases, for instance if labels contain blank spaces.

Graphing of data can also be accomplished by utilizing many of the tools used above to process the data. In particular the SciPy library offers a number of metrics for performing cluster analysis of the adjacency matrices produced above.37 Clustering methods are complicated and are an area of academic research in their own right, but various metrics can be suggested for best performance. In particular, “cophenetic correlation coefficient” is a metric that can evaluate the validity of a particular clustering technique together with a measure of distance or dissimilarity. Because we have already calculated the cosine similarity for each cell, we can simply subtract it from one to get the equivalent cosine distance, which can be used as the input for a clustering technique. When cosine distance is used to determine the closeness of data points, in this case cells from the Micro-zodiac, the best-performing clustering method was Unweighted Pair Group Method with Arithmetic Mean (UPGMA).38

36 Bastian et al. 2009, <https://www.aaai.org/ocs/index.php/ICWSM/09/paper/view/154> (accessed July 2, 2017). “Graphing programs” are tools that allow one to visually interpret numerical data and export the results as images. On csv file format, see in this volume, Martino and Martino, 137. For further information about Gephi, see in this volume, Pagé-Perron, 209.

37 <http://www.scipy.org/> (accessed July 14, 2017). For an example of performing cluster analysis to reveal objects typologies, see in this volume, Martino and Martino, 118–124, 133–134.

38 Saraçlı, Doğan, and Doğan 2013, <https://doi.org/10.1186/1029-242X-2013-203> (accessed July 2, 2017), 203, table 2.
there are many methods of clustering, the one chosen here is relatively simple in its procedure. It starts by looking for the two closest cells within the adjacency matrix and merging those into a cluster to which it assigns the average of their distance. It then looks for the next two closest cells, or, if the next cell is closest to the newly created cluster, it merges those and assigns it a new average distance. This process is completed for the entire adjacency matrix, merging close items whether they are cells or newly created clusters. When UPGMA clustering is applied to the adjacency matrix of cells from the Micro-zodiac, a dendrogram can be created which sorts the cells into hierarchical bunches based on their similarity.

This illustration shows the bottom part of the dendrogram, the cells which were easily clustered by the algorithm. The location of the cells is listed along the left side of the graph, and vertical connecting lines link cells or groups of cells into clusters. The vertical lines’ location along the bottom axis represents how similar the contained elements are. The left edge of the graph shows elements that are identical, and the right edge represents highly different elements. The graph clearly shows that cells with the same major and minor sign pairs are highly similar. The dotted line at .75 represents a subjective marker for where the clustering no longer seemed to produce relevant results. Each of these dendrograms encompasses the entire dataset regardless of the particular metrics of interest for the research. The .75 cutoff seemed to mark a point at which the clustering of cells or existing clusters was no longer based on shared ingredients but rather on average distances of existing clusters. The results point to a very strong association of cells with other cells sharing the same major and minor sign pairs regardless of the particular text. The bottom group of Gemini-Capricorn (3-10) come from three different tablets but all are identical in their medical ingredients. What is perhaps more telling is what clusters are not present. There are three clusters that have the major sign Cancer (4), each with a different minor sign, and they are all highly distinct. The inverse pattern is also true; there are clusters assigned to the minor signs Capricorn (10) and Aquarius (11), each with different major signs, and they are all highly distinct. If the association of ingredients was dependent on only the major or minor sign, we would expect the cells to cluster in a pattern in which either the major or minor sign were shared.

The same process can be applied to the inverse of the entire dataset. Instead of basing our analysis on the cells and the ingredients they contain, we can flip the paradigm around and look for ingredients that share similar cells. Here a similar picture plays out; the above graph is a detail from a dendrogram created using the exact same procedure outlined above (one of the great advantages is the rapid re-usability of methods on similar data). Naturally these clusters are slightly less organized because the dataset is considerably larger, and
connections are less common. Notice, however, that the ingredients cluster roughly into patterns that mirror the inclusion of ingredients in the individual cells: place, wood, plant, and stone. Interestingly stone is often slightly removed from the cluster of the first three ingredients. The likely cause of this pattern is the aforementioned common practice of including two stone ingredients in the cells, meaning that stone ingredients are more common throughout the entire dataset and thus slightly more distant from the other tightly connected ingredients.

Conclusions

The methods employed above took what was originally a damaged and fragmentary corpus of Late Babylonian astrological data and, through various quantitative processes, extracted meaningful information about its organization. While generally these methods are used in the analysis of much larger and more coherent datasets, the results here show that even text that is difficult to read in conventional ways still offers an opportunity for analysis with digital tools. The tools used here integrate well into the traditional philological
work of editing a text and offer another vantage point from which to gain insight about the patterns or structures perhaps not immediately obvious within a given corpus. The disjointed nature of the evidence is particularly well suited to quantitative methods that treat the fragmentary text as a series of data points and allow for the application of powerful algorithms. The patterns discoverable here are widely applicable to other corpora. For instance, one could use this same method to analyze the adjacency of the corpus of Neo-Assyrian letters to the king; texts clustered around each other might be a function of a single author or a particular topic. Similarly, a dendrogram could be constructed for the phrasing of royal inscriptions, illustrating the terms and forms shared among inscriptions by a single king over the course of his reign, or between or among kings.

As for the data derived from the Micro-zodiac texts, after the data for both the cells and the ingredients themselves were processed, the resulting graphs suggest that the medical ingredient cells of the Micro-zodiac are related to
each other only by the internal organization scheme inherent in the Micro-zodiac itself. If the cells are clustered according to shared ingredients, the most common clusters are oriented around cells sharing the exact same major-minor sign pairings. No other system of linkage or connection between ingredients appears in the clustered cells, and, even when the data is inverted to look for patterns between ingredients, no other organizational structure emerges. The results of the analysis show that the connections between medical ingredients in the Micro-zodiac text do not preserve an alternate system of association as might have been borrowed from another textual tradition. As mentioned in the introduction, there were a number of texts associating medical ingredients and zodiacal signs written during this period, and it is interesting that the Micro-zodiac does not seem to borrow from these.

The lack of a connection with contemporary texts or earlier forms of astrological knowledge suggest that the associations found within the ingredient row of the Micro-zodiac are a novel invention unique to this series. It is important to note that there are systems of association and connections that existed outside of the ingredient row, attesting to the Micro-zodiac’s borrowing of previous text and content. However, the lack of any explicit pattern in the ingredient row is important because it further illustrates the complexity of astrology during the late period in Babylonia. As more of the late astrological texts are studied, it is clear that the science of astrology was developing new genres of knowledge both by borrowing from previous traditions and inventing new forms of meaning.39 The Micro-zodiac texts were trying something distinctly new, but they were written in a context in which many other texts were also attempting new systems of organization or association.

References

Bastian, Mathieu, Sebastien Heymann, and Mathieu Jacomy. 2009. “Gephi: An Open Source Software for Exploring and Manipulating Networks.” Proceedings of the Third International Conference on Weblogs and Social Media, edited by Eytan Adar, Matthew Hurst, Tim Finin, Natalie Glance, Nicolas Nicolov, and Belle Tseng, 361–362. Menlo Park, CA: AAAI. <https://www.aaai.org/ocs/index.php/ICWSM/09/paper/view/154>.

Geller, Mark. 2010. Ancient Babylonian Medicine. Malden, MA: Wiley-Blackwell.

39 Koch 2015, 197–199.
Geller, Mark. 2014. *Melothesia in Babylonia: Medicine, Magic, and Astrology in the Ancient Near East*. Science, Technology, and Medicine in Ancient Cultures 2. Boston: De Gruyter.

Heeßel, Nils. 2005. “Stein, Pflanze und Holz. Ein neuer Text zur ‘medizinischen Astrologie.’” *Orientalia* 74: 1–22.

Heeßel, Nils. 2008. “Astrological Medicine in Babylonia.” In *Astro-Medicine: Astrology and Medicine, East and West*, edited by Anna Akasoy, Charles Burnett, and Ronit Yoeli-Tlalim, 1–16. Micrologus’ Library 25. Florence: SISME - Edizioni del Galluzzo.

Hunter, John D. 2007. “Matplotlib: A 2D Graphics Environment.” *CISE* 9: 90–95.

Jänicke, Stefan, Greta Franzini, Muhammad F. Cheema, and Gerik Scheuermann. 2015. “On Close and Distant Reading in Digital Humanities: A Survey and Future Challenges.” In *Proceedings of the Eurographics Conference on Visualization (EuroVis) (2015) STAR – State of The Art Report*, edited by Rita Borgo, Fabio Ganovelli, and Ivan Viola, 83–103. Aire-la-Ville, Switzerland: Eurographics Association.

Jänicke, Stefan, Greta Franzini, Muhammad F. Cheema, and Gerik Scheuermann. 2016. “Visual Text Analysis in Digital Humanities: Visual Text Analysis in Digital Humanities.” *Computer Graphics Forum* 36 (6): 1–25.

Jockers, Matthew L., and Ted Underwood. 2015. “Text-Mining the Humanities.” In *A New Companion to Digital Humanities*, edited by Susan Schreibman, Raymond G. Siemens, and John Unsworth, 291–306. Chichester: John Wiley & Sons.

Koch, Ulla. 2015. *Mesopotamian Divination Texts: Conversing with the Gods Sources from the First Millennium BCE*. Münster: Ugarit-Verlag.

McKinney, Wes. 2010. “Data Structures for Statistical Computing in Python.” In *Proceedings of the 9th Python in Science Conference, June 28–July 3 2010, Austin, TX (SciPy 2010)*, edited by Stéfan van der Walt, and Jarrod Millman, 51–56. <http://conference.scipy.org/proceedings/scipy2010/>.

Monroe, M. Willis. 2016. “Advice from the Stars: The Micro-Zodiac in Seleucid Babylonia.” PhD diss. Brown University.

Monroe, M. Willis. 2017. “Willismonroe/Monroe2017DH: Initial Upload.” Last modified July 14, 2017. Zenodo. <https://zenodo.org/record/827359>.

Moretti, Franco. 2007. *Graphs, Maps, Trees. Abstract Models for Literary History*. London-New York: Verso.

Pedregosa, Fabian, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, et al. 2011. “Scikit-Learn: Machine Learning in Python.” *Journal of Machine Learning Research* 12: 2825–2830. <http://jmlr.org/papers/v12/>.

Pérez, Fernando, and Brian E. Granger. 2007. “IPython: A System for Interactive Scientific Computing.” *CISE* 9 (3): 21–29. <https://www.computer.org/csdl/mags/cs/2007/03/index.html>.
Reiner, Erica. 2000. “Early Zodiologia and Related Matters.” In *Wisdom, Gods and Literature: Studies in Assyriology in Honour of W.G. Lambert*, edited by Andrew R. George, and Irving L. Finkel, 421–427. Winona Lake, IN: Eisenbrauns.

Rochberg-Halton, Francesca. 1988. “Elements of the Babylonian Contribution to Hellenistic Astrology.” *JAOS* 108 (1): 51–62.

Rochberg, Francesca. 2004. *The Heavenly Writing: Divination, Horoscopy, and Astronomy in Mesopotamian Culture*. Cambridge: Cambridge University Press.

Saraçlı, Sinan, Nurhan Doğan, and İsmet Doğan. 2013. “Comparison of Hierarchical Cluster Analysis Methods by Cophenetic Correlation.” *Journal of Inequalities and Applications* 2013: 203. <https://doi.org/10.1186/1029-242X-2013-203>.

Sinclair, Stéfan, and Geoffrey Rockwell. 2015. “Text Analysis and Visualization: Making Meaning Count.” In *A New Companion to Digital Humanities*, edited by Susan Schreibman, Raymond G. Siemens, and John Unsworth, 274–290. Chichester: John Wiley & Sons.

Steele, John M. 2008. *A Brief Introduction to Astronomy in the Middle East*. London: SAQI.

Steele, John M. “Astronomy and Culture in Late Babylonian Uruk.” 2011. In “Oxford IX” *International Symposium on Archaeoastronomy Proceedings IAU Symposium* 278, edited by Clive L.N. Ruggles, 331–341. Cambridge: Cambridge University Press.

Steele, John M. 2015. “A Late Babylonian Compendium of Calendrical and Stellar Astrology.” *JCS* 67: 187–215.

Walt, Stéfan van der, S. Chris Colbert, and Gaël Varoquaux. 2011. “The NumPy Array: A Structure for Efficient Numerical Computation.” *CISE* 13 (2): 22–30.

Waskom, Michael, Olga Botvinnik, Drew O’Kane, Paul Hobson, David C. Gemperline, Yaroslav Halchenko, Saulius Lukauskas, 2016. “Seaborn: v0.7.1 (June 2016).” *Zenodo*. Last modified June 6, 2016. <https://zenodo.org/record/54844>.

Weidner, Ernst F. 1967. *Gestirn-Darstellungen auf babylonischen Tontafeln*. Vienna: Böhlau in Kommission.

Zundert, Joris J. van. 2015 “Screwmeneutics and Hermenumericals.” In *A New Companion to Digital Humanities*, edited by Susan Schreibman, Raymond G. Siemens, and John Unsworth, 331–347. Chichester: John Wiley & Sons.
PART 4

Online Publishing, Digital Archiving, and Preservation