Jonathan Armoza

Reading and Writing Archives

Tom Augst

December 23, 2016

**Model as Archive: Computational Perspectives on Emily Dickinson**

**Introduction**

Today, let’s go on a journey across an archive. To curate a collection of objects is to create a narrative about them. Though an “archivist” has collected objects into one container, perhaps described them with criteria of their choosing, and then presented them to an observer in a particular way does not necessitate a relinquishment of observational authority. Rather, coming to an archive and “looking” infers participation in the meaning-making of that archive. The archivist has given an outsider a window into the systems of meaning created by those objects and a juxtapositional system of meaning to represent them with. If meaning-making via an archive was not participatory, the intent of preservation would be undone. Of course, the archivist cannot anticipate how meaning-making will occur at the site of interface with their work, but they can provide tools of access to do so. Tools of access preserve contemporary ways of meaning-making as much as they do to provide access. Archive creation with an equitable impetus toward curation and representation is no different than when a mathematician, engineer, or scientist creates a model of a data set or more theoretical objects.

Models of data are themselves systems of objects and inevitably representations of physical objects and physical phenomena. They reflect back at the physical world narratives of representation, and in separate instantiations they do so in unique, linear ways. Think of the shape of a curve in a relational, two-dimensional graph which poses one quality against another. The slope of the line infers a relationship dynamic and predominance of left to right graphing infers a space for those dynamics to order themselves. Space-time is our preference. It can be difficult to remember just much it is a preference. The past is past, the future is future, and the present is so ephemeral it can hardly be mentioned. Past and future are often rhetorically situated to justify a present quality. Linear order, after all, connotes and denotes an argument. It is a juxtaposition imposed by the components of our physical reality (e.g. four dimensional spatial and chronlogical existence) and representations of those phenomena (e.g. visual representations of those orderings). Every time we create an order, we instantiate a narrative. Like a model, the order in which objects are presented in an archive matter a great deal to the meaning-making of that archive.

And since meaning-making is participatory, granting the observer of a model or archive, the tools to derive an order relevant to their own perspectives is paramount for either kind of collection. To be clear, archives and data models share a performative motive for exploration. A data model of an archive – a tool of access – can open up the possibilities of archival exploration. Such a model would itself preserve (or “archive,” if you prefer) those exploratory possibilities and, of further import, the flaws of the systems of representation – both the archive’s and the model’s.

**Shape**

There is, indeed, *no* *time* like the present. As much existential investment and weight is given to chronological order, it should not be surprising to a cultural researcher or scientist how much of an interest we have in exploding the hegemony of that order, and reconstructing and re-layering it in a way that is meaningful to our present. One need only look at the description of a demarcated selection of time: time*frame*. The character of that demarcation is contentious. The danger of anachronism in a chronological context is the possibility of a misjudgment about a phenomena within that context.

Carolyn Dinshaw offers an anecdote on anachronism in the beginning of a chapter entitled, “Out of Sync in the Catskills,” that illustrates this point well. Describing an area prone to flooding in upstate New York where she once lived, she tells of how that flooding provided new insight on the metaphor of the “stream of time” – something that flows “unidirectional and evenly, producing a concept of time as smoothly progressive, one moment coming after the last in a steady flow” (Dinshaw 129). Of course, as she notes, “streams do not always behave this way” (130). Observing a nineteenth century gravestone wash ashore on the banks of the creek near her home, carried downstream from a local cemetary, it strikes Dinshaw that the time she is observing is “heterogenous and always already very full” (130). The asynchronicity of this timeframe provides a perspective unavailable in the spatial linearity of the cemetary to her creek bank, unavailable in the chronological linearity of that nineteenth century burial to her twenty-first century flood:

“The ghastly rock provides a powerful figure of asynchrony, realized in a particular moment and in a particular place but also opening up a spectral view of the world, a perspective in which the boundaries between the living and the dead, the material and the immaterial, the real and the fictional, the present and the past are porous. It is a world in which everyday time is itself experienced as wondrous, and the present, unmasterable, is full of other times” (130).

This sort of asynchrony extends beyond the physical world to our representations of it, and the multiple systems we construct to narrativize those representations. The book or “text” as object looms large in the history of knowledge production for such representations. It is itself a system, the purpose of which is declared, and inherently, vaguely so via its symbols which are meant to encode meaning(s). Consider for this writing a system as a collection of things set about to convey or explain a whole. Quite ironically under this definition, the contents of a text need not convey a semantic completeness or conclusiveness as the reader might wish. The words, punctuation, and sentences are put into place to convey some thing. The text is also a representation that despite the linearity of its narrative, which Augustine utilized, frustrates how physical time operates – in the sense of the ticks of an atomic clock, for instance. Texts do progress with a “text-time,” a time-sense delivered much like the presumed reading experience of taking in one word and sentence after another. It is “presumed” because this is not often reflective of the reading experience at all. One takes a look backward, or sometimes even ahead in a text. In physical time, as we currently experience it, one cannot reify the past by actually reliving it or foreshadow the future by skipping ahead a few days only to return to the present.

This is also part of the reason why the metaphor of text as a readily visible system of things is such a successful analogy. Each unit or *form* that is the text or is within the text is set about to convey a meaning. Writing of system and the “natural selection” performed by genres of knowledge, Clifford Siskin writes:

“As an alternative to the history of ideas, the history of mediation tells us to look to form as well as to content, and the history of the real, although a necessarily blunt instrument, scales up our perspective on how forms change and how change functions over time” (Siskin “Coda”).

The objects conveyed within a system of meanings are meant to be layered and considered in that arrangement. In other words, form necessarily shapes and mediates content. Each successive sentence, formula, observation, notion in a book is built upon each prior one conveyed, in addition to all of the sentences, formulae, observations and notions that it takes to understand them. At the interface of text to reader, we see how encoded messages of meaning making require participation of the observer. They also require an array of asynchronously posed, theoretical-semantic objects. This meaning-making activity that is the center of the outcomes I have mentioned thus far – the headstone washed ashore and book-as-system – is the act of decomposition.

Andrew Piper writes of this activity in the context of literary study and data models of literature in “Reading’s Refrain: From Bibliography to Toplogy.” Looking to Augustine’s concept of the “conversion” of a reader that linearly experiences the ideas conveyed in a textual system like a book, Piper asks us to reconsider the historical and experiential weight of nostalgia we have for that reading experience. He looks to Foucault’s idea of dispersion of texts as a way of understanding latent relationships that exist among and within them. To think of the words of a text as the objects of an archive, why might we insist upon a static and deadening fixed order for those words? Here we approach, very briefly, closely held notions of authority. The fixed order of a text is what denotes an imagined, unmasterable and unmatchable positioning of meaning making. Reconsidering that order forces us to ask very uncomfortable questions in that they might, if temporarily, violate the act of archiving. For instance, what would a resultant textual object be after a transformation that removed an arbitrary amount of its sentences? What if one randomly re-ordered a text’s paragraphs? Questions like these all lead us to the same question: What does the decomposition of an archival object do to its intended meaning-making? In the case of a text, the argument that an arbitrary model makes for us is that decomposition of meaningful units of it can bring us a more comprehensive understanding of how those units are employed for that meaning-making. This language of decomposition and consideration of proportional relation is inherently mathematical.

In “Reading’s Refrain,” Piper aims to demonstrate a lineage between this new kind of text analysis that sets textual elements into quantified relation – what he calls "topology" – and its humanities-theoretical predecessors, most notably structuralism and poststructuralism. He articulates in humanities terms how the quantification of text operates and how and where it must borrow from across the supposed epistemological divide between the alphabetic and numerical: "[Topology] is an attempt to reverse the false notion, in Friedrich Kittler’s argument, of the two-thousand-year- old antipathy between the alphabetic and the numerical" (“Reading’s Refrain: From Bibliography to Toplogy” 380). To bridge the chasm that is necessary for this decomposition, we supply text analysis with the mathematical notions of cardinality and continuous space. He continues, the "recourse to number — however inelegant for many — is an attempt to move past the ontology of discourse," moving us toward the claim that "for every unity [e.g. your choice textual/cultural artifact of study] there can always be something between it and that which it succeeds" (381). What is countable? And what is not? What is categorizable? And what is not? These notions of countability and discreteness and the proportionality that they imply are not new for the humanities. Discrete and expected sensibilities abound from acts of reading, semantic ones like polysemy and betweenness. As Piper wrote in his *Dreaming in Books* four years earlier:

"Armed with the tools and training of bibliographers and book historians, numerous literary scholars today have begun to draw our attention to the ways that bibliographic details are key determinants, but also key multipliers, of textual meaning" (*Dreaming in Books* 9).

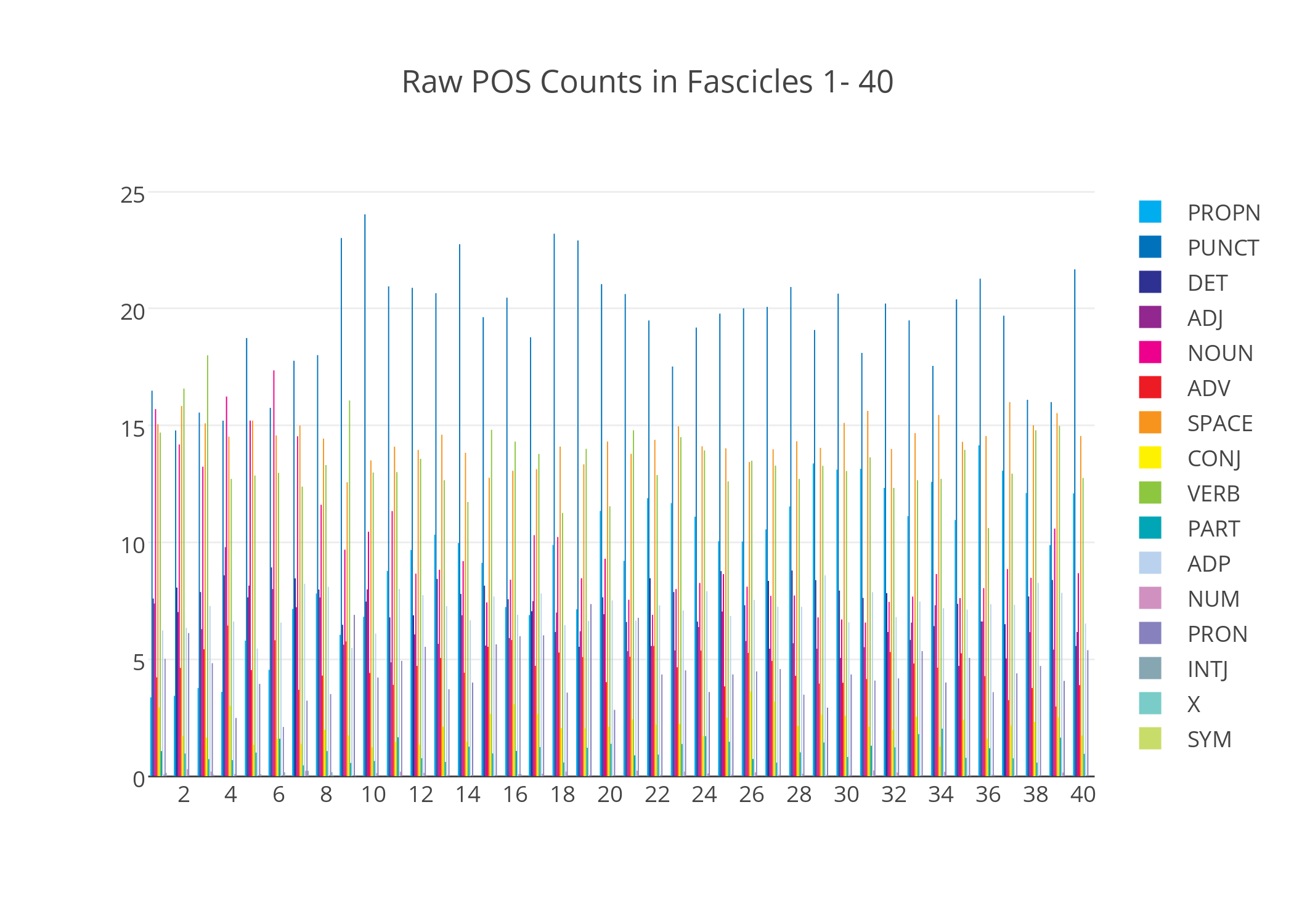
Ripping apart a text to count it and visualize it in varying ways produces another kind of conversion: new understandings of the lexical pieces of the text in relation to one another and how those relations play out against extra-diegetic or cultural contexts. The gravestone of the text washes ashore in a new, twenty-first century frame: the computational model. Whereas archive/text offers one possible representation of its contents, model as archive exponentially increases possible representations of them. With this new tool of access that enables these new understandings, the expectation of reading a text as a linear meaning-making of experience from a static object falls away with its components explicitly split apart. But it does not go away entirely. That reading can but does not always progress linearly is already established by older forms of knowledge like the book, and those forms still mediate knowledge if but because it is in that state in which they were initially presented. This contradicts the idea that computational “dispersion” or “decomposition” of a text is a new tactic for examining it. Rather, it is the computational algorithm rendering a model of another system that is new. What is more conscious about the use of the computational algorithm to mediate content is the sense of experiential process and form that it makes visible. Even the algorithm itself is a “process written up in a certain form,” shaped by the fundamentals of the discipline of computer science and the technology (hardware, operating system, programming language, and compiler/parser) upon which it it relies (Siskin “Coda”). Its steps are often available to its users in the form of source code, the execution of which is exposed in the user-controlled act of debugging.

**Model**

When Emily Dickinson died in 1886, she left behind a body of unpublished writing significant in size and form. Of the roughly 1,800 individual poems, over 800 were handsewn into forty manuscript books called "fascicles." The page orderings of those books were lost as a result of a family feud over ownership rights and differing ideas about how to organize and publish such a large collection of poems. However, those page orderings were recovered by the forensic analysis of R.W. Franklin and others, as initially documented in *The Manuscript Books of Emily Dickinson* (Dickinson 1981). Still, the size of the collection has made comparative readings of the poems a difficult task. The fascicle orderings of Dickinson’s poems have become for some scholars a structuring basis for uncovering thematics or narratives in her writings.

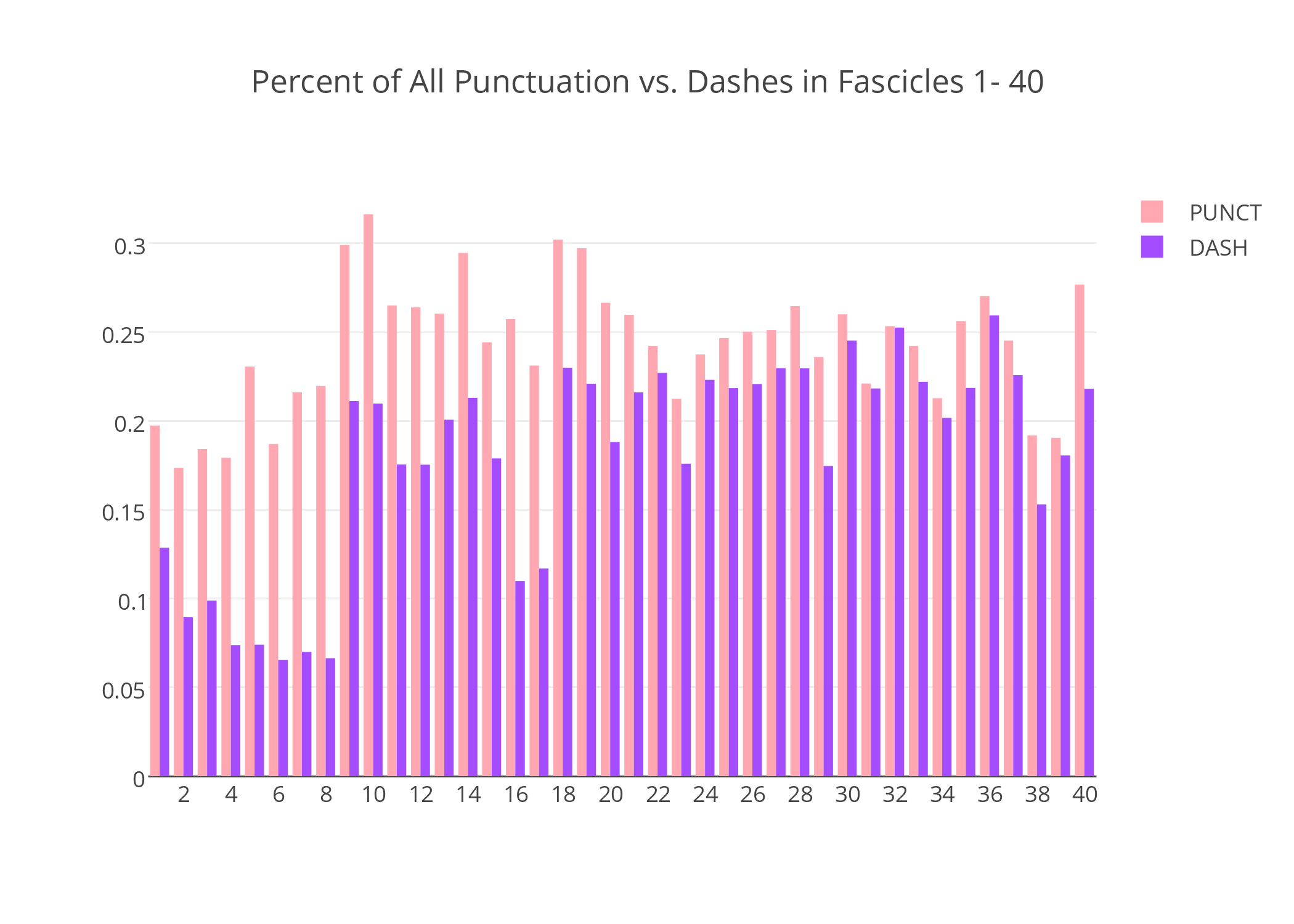
Thus far we have considered the idea of a computational model in a very general sense. One way of thinking about the decomposition of an archive is to count its features for comparative purposes. In the case of a textual archive, we plainly have its words to count. For instance in those roughly 1,800 poems there are, from a cursory accounting, 20,217 unique words, and in those 800+ fascicle poems 12,143 unique words. That is not all there is to count. Words and the ways in which they are employed denote relationships. In this light, the amount of potential relationships between word usages and the complexity that amount imbues quickly outpaces the capacity of human consciousness. There are many ways to break down this complexity into digestible and meaningful readerly bites. These are statistics quite easily included alongside or over top an archive, and can radically reconfigure the understanding and exploration of it. For instance, we could look at the parts of speech (POS, singular or plural) employed within it and provide an account of those usages to facilitate its exploration.

Thanks to the efforts of researchers in computational linguistics, natural language processing, and machine learning we now have fairly competent tools for determining the POS to which each word in a text belongs. Such “tagging” of POS relies on word usage context as well as the “experience” (i.e. a previous model) of other same-language data sets in which POS are considered to be known. One recent text parsing API that tags POS is “spaCy” by Mark Johnson and Matthew Honnibal.[[1]](#footnote-1) We will employ this POS tagging over the Dickinson fascicle corpus, accounting for proper nouns, punctuation, determinants, adjectives, nouns, adverbs, spaces, conjugations, verbs, participles, adpositions, numbers, pronouns, interjections, and symbolic characters. All remaining untaggable words are accounted for in a category, X. The words of each text are tagged with a POS. Counts of those tags are vectorized (turned into lists of numbers) to represent each text. Texts of varying size thus have summed counts for each POS category across the syntactical patterns of POS they contain, word order being disregarded after the tagging occurs. Looking at raw sums of those sixteen POS categories across the forty fascicles – which are numbered in Franklin’s arrangement in a presumed chronological order of their construction – we get sixteen separate dynamics brought into juxtaposition in the image below.



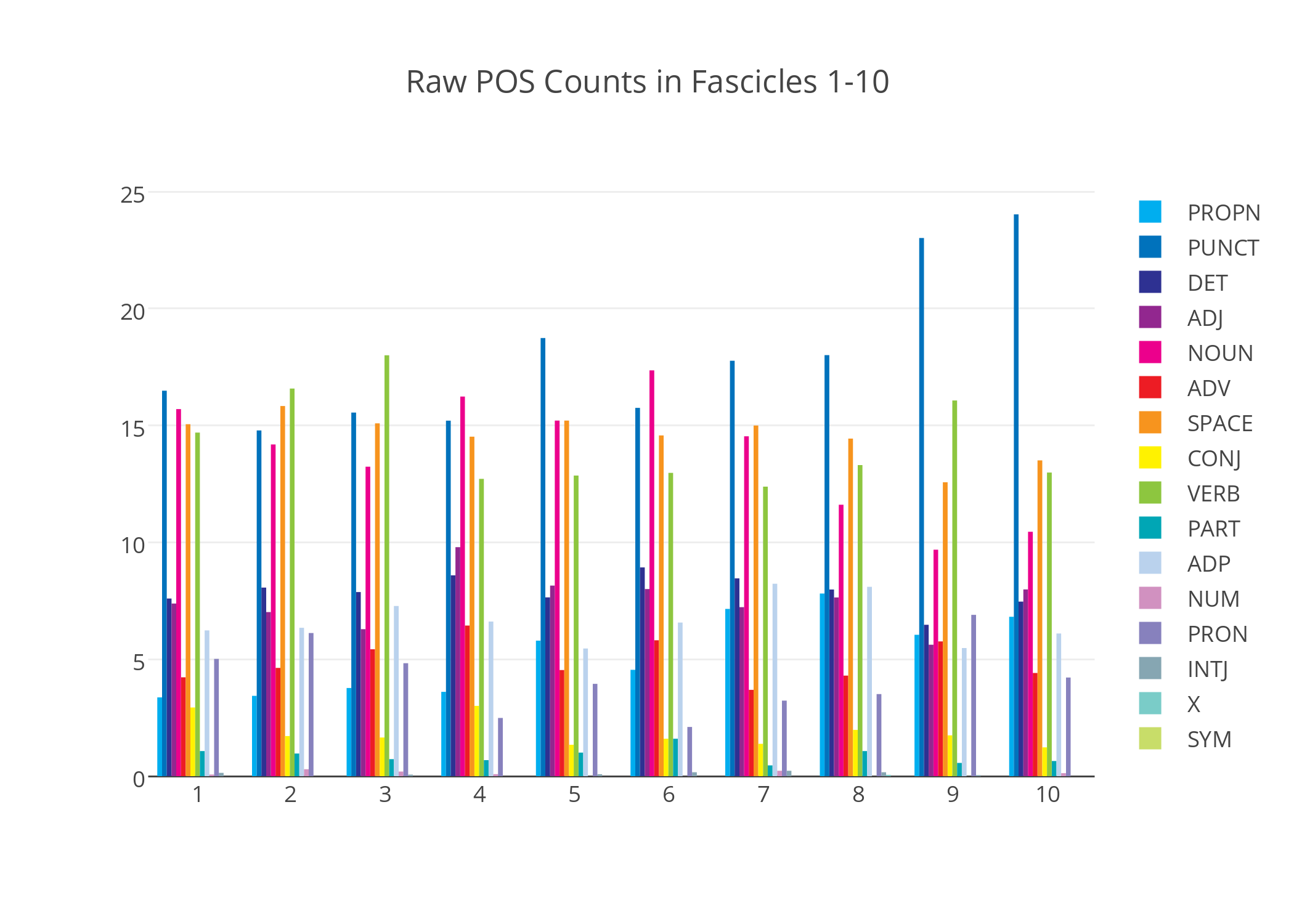
**Figure 1.** Raw counts of the POS found in each of the forty fascicles as tagged by spaCy.

The view quickly becomes cluttered as we try to visualize sixteen elements of a system of forty things. We could further isolate those counts to see them more clearly, but there are also several trends available from this image. Punctuation, in dark blue, emerges as the most utilized POS only after fascicle eight.[[2]](#footnote-2) It also diminishes in fascicles 38 and 39. This in itself is a curious phenomenon, though if one considers Dickinson’s notable use of the dash, it might not be so surprising.



**Figure 2.** For all forty fascicles, dash counts (as percentage of overall POS) are set alongside all punctuation (also as percentage of overall POS).

An examination confirms this assumption, and in remarkable proportion worthy of a full, separate study. The introduction of dashes significantly increase with fascicle 9, and this adds credibility to Franklin’s work (and Thomas Johnson’s, ca. 1955) that fascicle construction also roughly corresponds chronologically with the authorship of the poems they contain. Dashes also drop to pre-fascicle 9 levels in 16 and 17. What is of further note though is why that punctuation (and/or the confirmed dashes) are in relatively lesser proportion to other parts of speech in the first group. To think of the basic model of this archive of texts in a mathematical sense, we see that the parts of speech of texts are considered in sixteen dimensions, each of which may provide a component of influence to alter the overall style of each fascicle. Let’s zoom in and take a look at fascicles 1-10 where punctuation is comparatively lower.



**Figure 3.** Raw counts of the POS found in each of the first ten fascicles as tagged by spaCy.

Glancing between Figures 1 and 3 we see that these initial fascicles are notably verb (green) and noun (pink) heavy in comparison to the latter thirty books. Nouns and verbs have shifted places as superlative POS features of these poems, marking a now quite evident shift in Dickinson’s style. Might this mean that she shifted from action heavy presences to some sort of absence with the introduction of the dash? This ground has been anecdotally covered by past Dickinson scholarship. The computed model offers a confirmation or extension and chronological demarcation of that premise. Taking averages of vectors though (already a reduction of the complexity of the original text) effectively diminishes the individual POS dynamics that were so carefully mined from each text. Close reading can alleviate this, but depending on the numer of poems in each fascicle (up to forty in some) we return to anecdotal exploration of a sizable number of objects in the archive.

It is important to remember that in this kind of examination above we are looking at counts from a probabilistic algorithm. In a sense, the foundation of the POS model is hidden from us if we just look at these counts without consideration of the method by which they were derived. The POS tagger for "spaCy" is a described as an "averaged perceptron tagger...with Brown cluster features" that utilizes ~1GB of English language training data and performs with a claimed 93% accuracy (“Citation Information #272”). This means that both language external to the corpus was considered in the act of POS tagging, and context based analysis of the words of Dickinson’s poems were assigned to clusters of words with high likelihood of them belonging to the same POS. And those determinations were averaged over a set of examinations of those words. So the tagging is approximate and has an expected percentage of error. This less reflects the weaknesses of the tagging method, and more the complexity of the potential relationships between the words of a text. But this too is not a new area for disagreement. While inherently more informed decisions, the assignments of POS to particular words by human experts do not always necessarily concur.

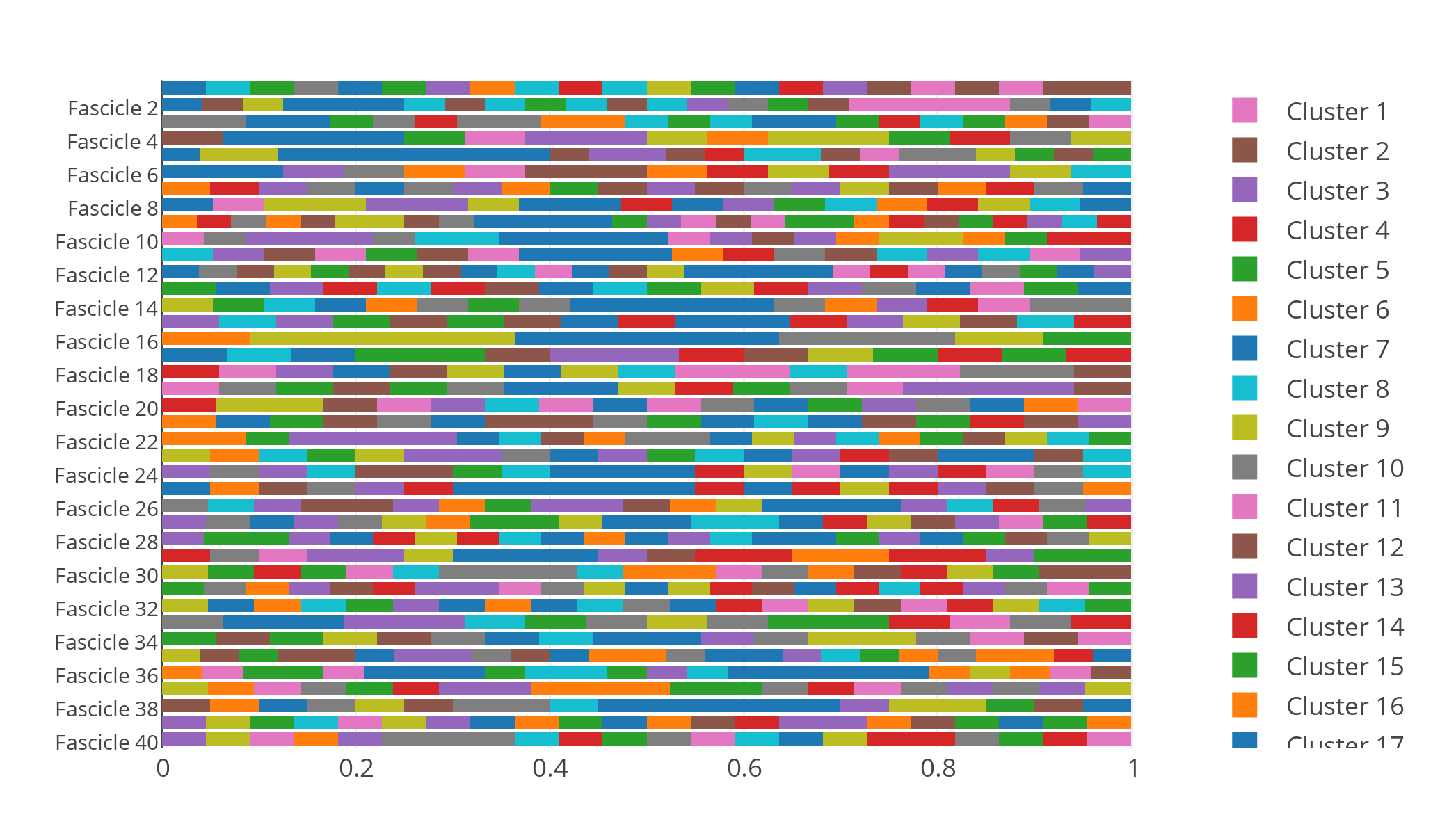
Making sense of a sixteen dimensional problem let alone a 12,143 dimensional one is not a task to be taken lightly. How can we reduce the dimensionality of these problems to ones that are more sensible for our purposes of traversing an archive? While we can divide and conquer the results of computational modeling for analysis purposes, we can also model the model itself. This is the next level of understanding of the things of the world around us as asynchronous systems, and it is the basis from which most Internet-based technologies with billions of users and millions of websites now function. With increasing complexity in an ever-expanding archive, if we are to allow for that room for probabilistic error, insightful understanding of its contents via these new forms is concomitant. One dominant area of research into these probabilistic models is considered to be Bayesian in philosophy and method. Utilizing the Bayesian notion of conditional probability and prior information, we attempt to determine the probability of latent factors affecting our data given that data. Such prior information encodes our assumptions about a data set and is added to an overall model of that conditional system (with its multiple elements varying over categories or values). Once computation of that conditional probability is made, we can come to conclusions about that data through an outputted “posterior” distribution of probabilities that reflects those latent factors (a.k.a. “latent variables” or “hidden elements”). This provides us with concrete descriptions/categorizations of the data. Those hidden elements may be labelled by us as reflecting familiar characteristics of a data set – like “topics” for text data – but they are none the less “things” that are reflective of now-modeled consistencies within that system of observed data.

One recently popular set of methods for identifying clusters of vector data is known as collaborative filtering. These methods are used in several common contexts such as recommendation systems. Netflix, for instance, sponsored a collaborative filtering contest in 2006 that challenged researchers to make improvements above their algorithms’ accuracy for inferring user ratings of movies. The general premise is that between its users and its movies there are attributes of either that draw one to the other and elicit a response. What are the basic questions that they would be interested in answering via model of their data? (1) What about a user makes them watch a movie and/or give it a particular rating? (2) What about a movie makes users watch and/or give those ratings? It is suggested that the answers to those questions may be characterized or approximated by recovering those latent factors. Note that the factors are not the answers, but they are suggestive of them. This is the power and weakness of presenting a system in this way. The model encodes our assumptions about a problem set and proceeds to approximate answers to questions we might have, given what we can observe. If done correctly, once the result is determined, an assessment is made over that result and the model is revised according to external information (other data or our intuition).

**Decomposition**

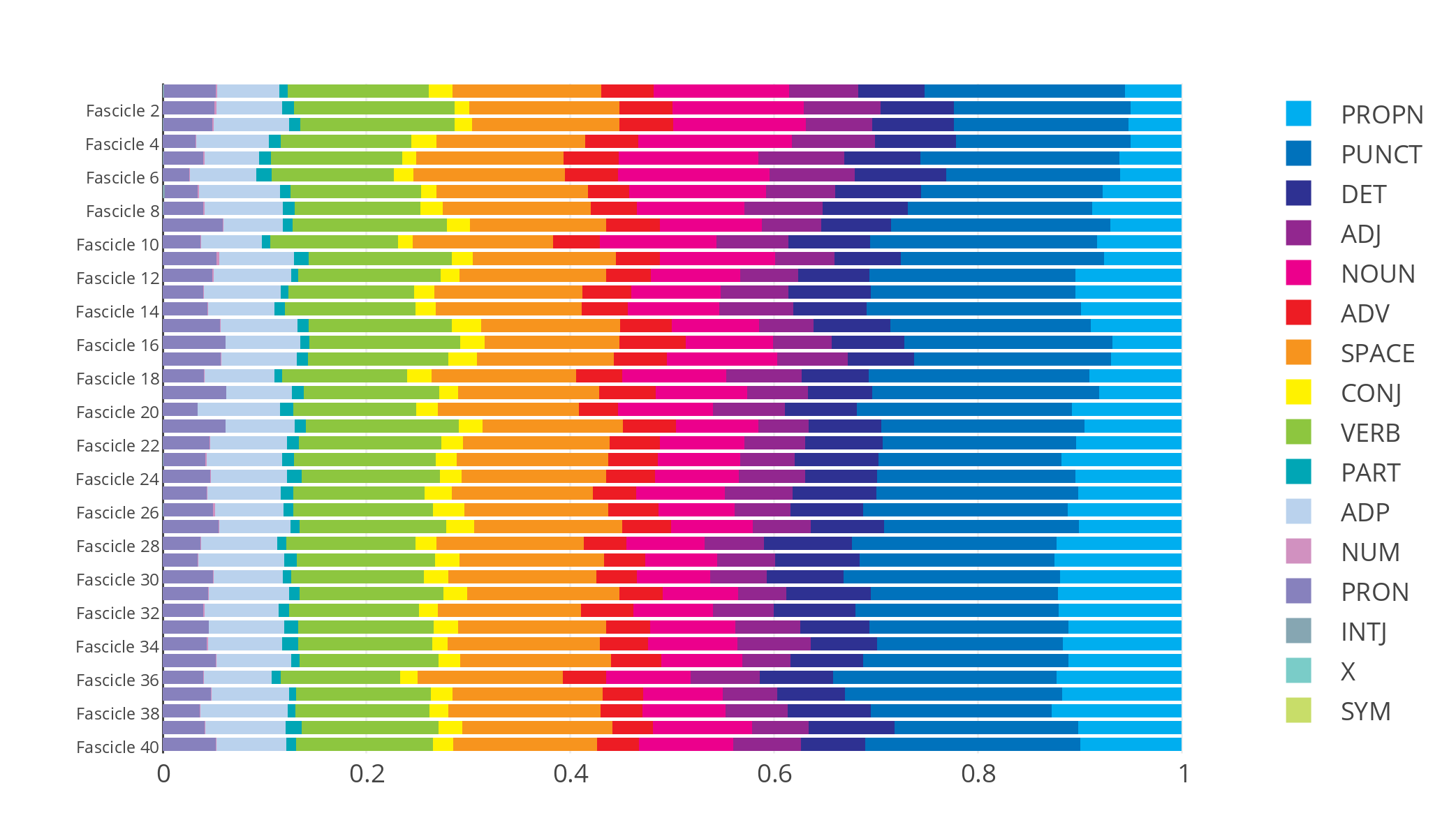
One of the successful methods of collaborative filtering is – in the very same sense – decompositional: matrix factorization. Here it will be enough to lightly describe the mathematics involved in this process. Each POS vector (recall, a list of numbers) represents a linear equation with a set of known coefficients (the POS category counts). A vector with three coefficients could be written as such: . Given a set of vectors we stack them up into a table. In linear algebra this stack of vectors is referred to as a matrix. In formal notation we erase the variable names for clarity. Let’s say we had three POS vectors, representing three texts. A matrix would look like a row by column table of those coefficients. In the Netflix examples these counts are replaced by ratings, each vector representing a user. And there may be large holes (zeroes) in this matrix due to missing values. One task could be to infer what those missing ratings might be based on the ratings of other users who have watched similar movies to a user of interest. Another task could be simply to group users for some recommendation purpose. The latter is what matrix factorization is more typically used for today.[[3]](#footnote-3) In either case, we would like to know how those latent factors contributed to the values of this matrix and to do so we factorize it, attributing some portion of each of its values to latent user factors and the remaining portion to latent movie factors. This produces two matrices, one for users and the other for movies, the coefficients of which will inform the proportional contributions to our known data: the stack of user rating vectors.[[4]](#footnote-4) The idea is that these two matrices can be multiplied together to produce our observed data – thus they are factors. The values of these matrices are determined over a number of iterations of the factorization method until an approximated error value is deemed stable enough (i.e. factorization results consistently produce an error value of x%). For the purposes of this writing, we will look to a Bayesian approach to this method called Probabilistic Matrix Factorization (PMF). Matrix factorization, in general, solves the problem of high dimensional data. One implicit but important property of this decomposition is that in order for matrices to be multiplied they must share a dimension. If one has a height of 3 users, the other has to have a width of 3 ratings, and so forth. This has the effect of allowing us to take a matrix of 50 million users and 50,000 movies and reduce one dimension each of the resultant factor matrices. The suggestion of a lower dimension (it could be 1,000 or 2 or whatever we feel proper) is given to the method and this is the height of one matrix it approximates and the width of the other matrix it approximates.

The Netflix example helps our understanding of this method, but users and items are only one metaphor for this kind of model. Another metaphor more appropriate to our POS-tagged data set is one proposed by biologists: patients and genes. In this modeling metaphor, we try to identify latent factors that are “hidden patients” and “hidden genes.”[[5]](#footnote-5) The task is then to identify clusters of patients with similar genetic traits, or the inverse, genetic traits with similar patients. We overlay this metaphor to the text to POS problem and can produce “hidden poems” or “hidden POS.” In this case, we look to identify those hidden poems around which we can group poems as somehow linked by similar POS usage.[[6]](#footnote-6) We run the factorization algorithm several times to produce a consensus of the coefficients of those matrices, determine a hierarchy of those consensus results and then flatten the hierarchy to produce a cluster assignment for each poem. These results tell us much more than a mere cluster assignment ID though. In fact, they present an association of each poem with one of those hidden poems which are proportionally representative each cluster poem’s features, in this case, their POS counts. They are also far more representative of the collection of objects as a whole. This is the model of the model. It is system algorithmically produced. We will refer to these representations as PMF produced POS profiles, or PMF-POS profiles for short. They enable some vastly different understandings of a collection of objects, that are simultaneously more comprehensive and individually representative. For instance, let’s substitute these profiles into each fascicle as representatives of their poems, assign profiles a unique color, and chart them. Below we list profile presences in proportion (out of 100%) for each fascicle with each book listed from top to bottom:



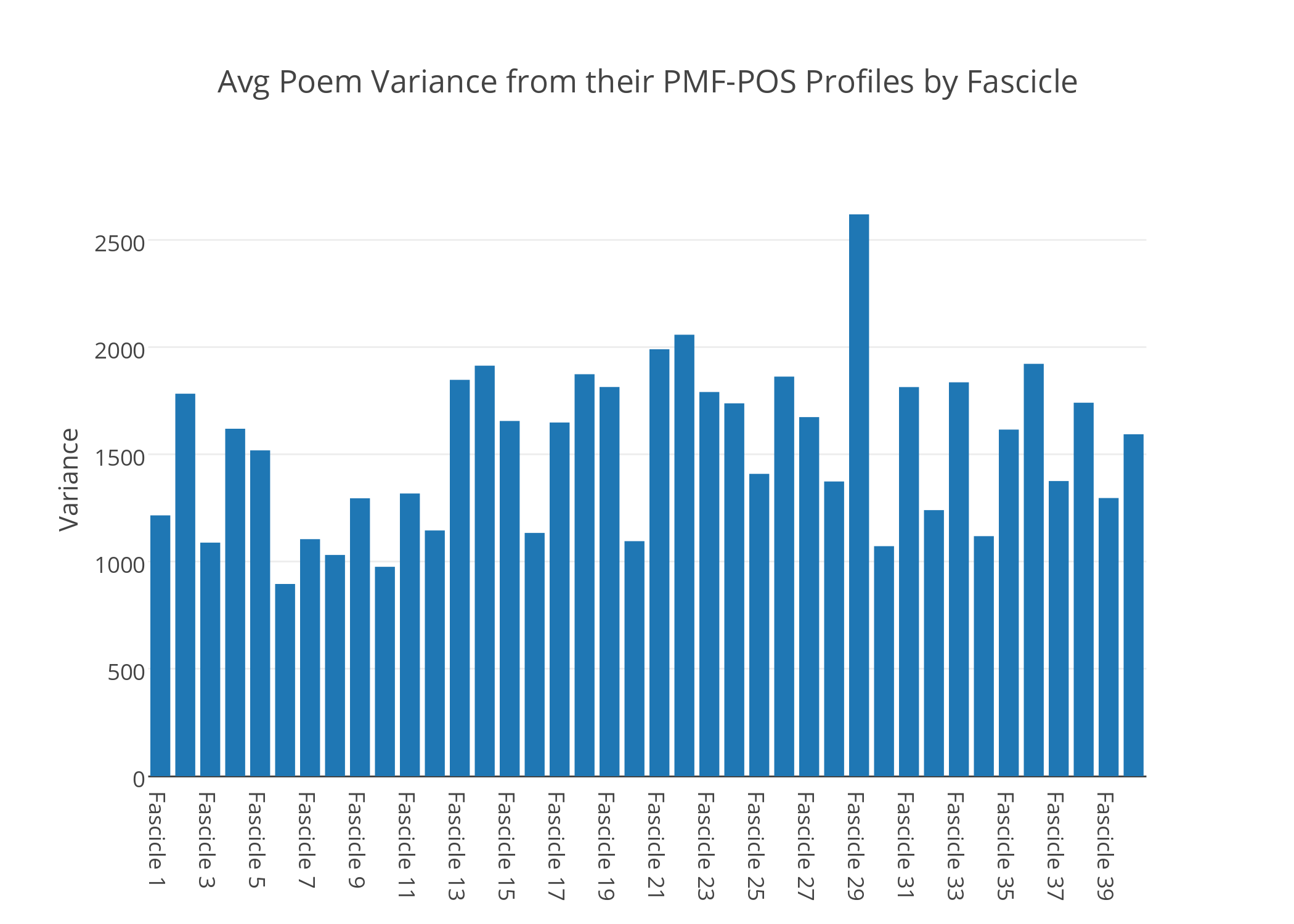
**Figure 4.** Proportional PMF-POS Profiles of Emily Dickinson’s 40 Fascicles

Like with Figure 1, this information quickly becomes unreadable due to the sheer heterogeneity of the 247 profiles found by PMF modeling (and those 247 color values cannot fit here for space reasons). Though profile/cluster 7’s POS pattern does seem to be be present in a significant set of fascicles we can see that diversity of patterns of POS speech usage is evident in Dickinson’s arrangements. Perhaps we can get a clearer view of this information in another way to present these profiles in a dynamic per book. To further refine the PMF-POS profile we calculate some more approximating information about them. We average the POS vectors of each poem identified as being assigned to a cluster, and then given that mean profile vector we determine the within-cluster variance of each poem’s POS vector from that mean vector (Halkidi). Though we don’t know the exact shape of the statistical distribution from which these POS vectors might arise (we can guess via the statistical family of prior distribution given to the modeler), we do now have approximate means and variances from such a distribution. We can use these values to navigate this model of a model of an archive. Below we reconfigure the view in Figure 4 to show the average POS of the mean profile vectors of the poems in each fascicle. (Each POS is assigned a unique color, and again shown in proportion out of 100%).



**Figure 5.** Average POS of the PMF-POS profiles used in Emily Dickinson’s 40 fascicles

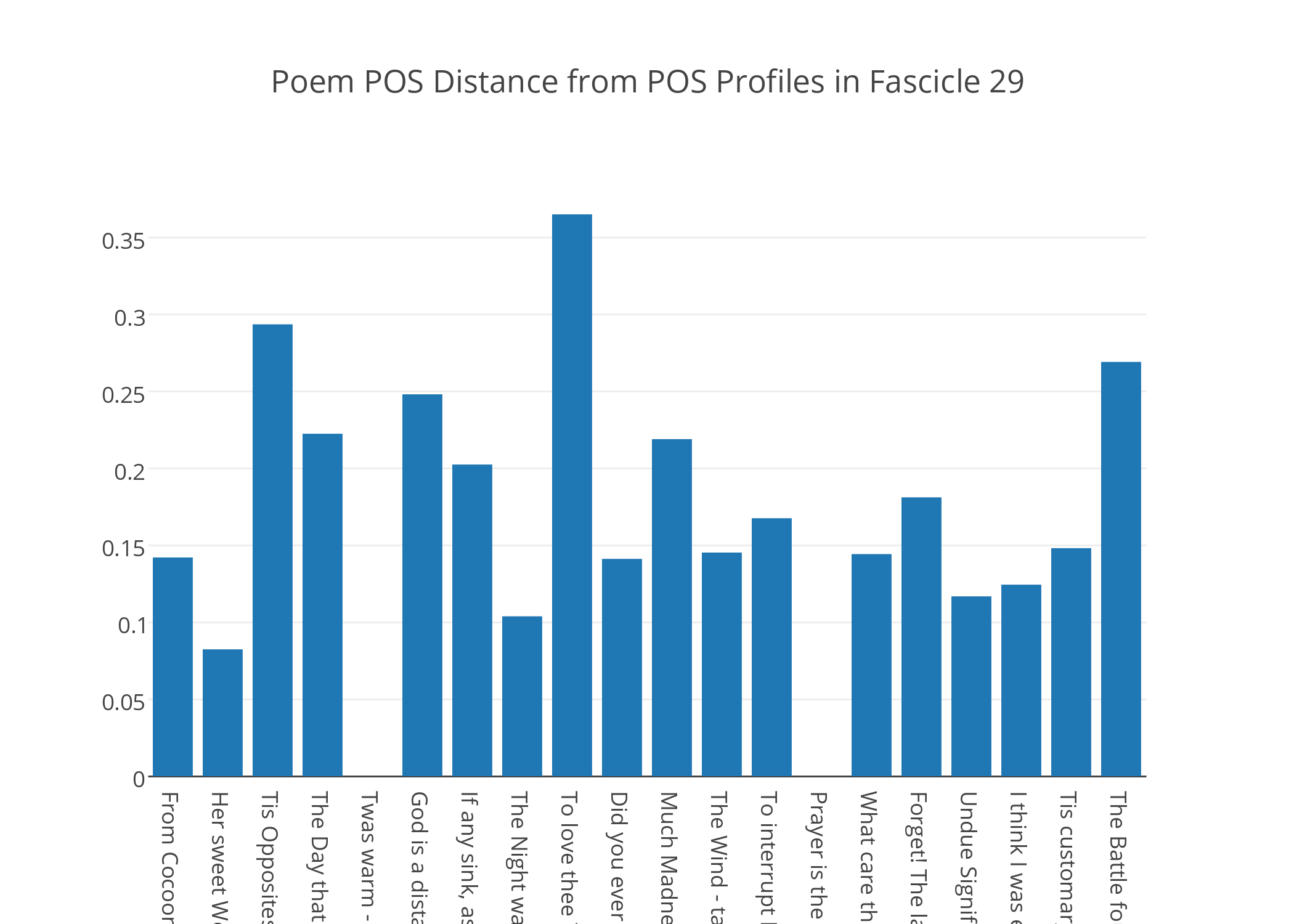
Now a clearer dynamic emerges of how POS operates with respect to this model throughout the fascicle books. We could split this view up to more easily identify those separate POS dynamics (and we will below), but we can see that some of the more pronounced POS trends of Figure 1 have been tempered. Punctuation, for instance, remains a large presence in the corpus, and proper noun usage does spike after the first few fascicles, but now the growth of each appears somewhat more gradual. What follows below is a more directed examination of this PMF-POS profile model of Emily Dickinson’s fascicles. I offer the reader the question of what would happen if this kind of model were provided alongside easier statistics like word counts and parts of speech. What might an exploration guided by this model look like? A final visual of these PMF-POS profiles below offers a guide. The variance of poems from their model profiles provide the first key step in model exploration.

****

**Figure 6.** The average variance of poems from their PMF-POS profile across fascicles

**Exploration**

It is quite puzzling this dynamic of variance, but statistically and metaphorically it is where exploration begins. The interests of mathematics and arts push against each other here, the former typically concerned with providing general explanation for phenomena and the latter typically interested in standouts. There is a sensibility that says that if variability from a norm of an observation is large, we have poorly classified it. Another compatible sensibility demands a look at it to see why it varies. What is that feature that stands out on the horizon? The question I am wondering about in Figure 6 is, “What is it about Fascicle 29 that makes its poems so different than their POS profiles?” Let’s a take a look at this fascicle to see if we can find the answer to that question.



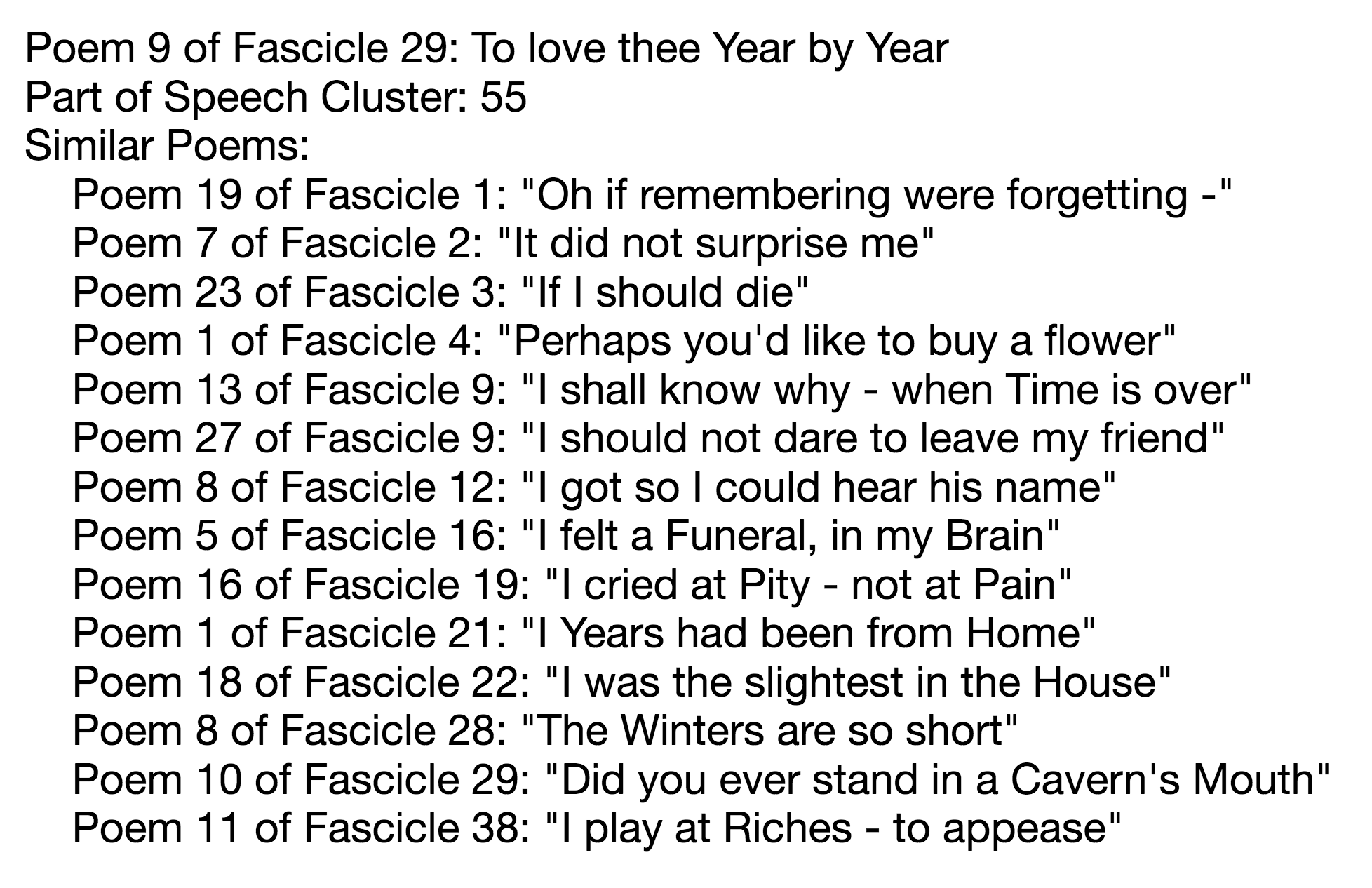
**Figure 7.** Jensen-Shannon distances between the poem POS vectors of Fascicle 29 and their respective PMF-POS profile vectors

Of the twenty poems in Fascicle 29, the POS vectors of two poems, “Twas warm – at first – like Us” and “Prayer is the little implement” precisely match their PMF-POS profiles diminishing the average variance that drew our attention to this fascicle. However, three poems, “To love thee Year by Year”, “The Opposites ­– entice“, and “The Battle fought between the Soul”, exceed the average distance of poem POS vectors from their profiles by more than 25%.[[7]](#footnote-7) “To love thee Year by Year” is a short poem that exceeds this average distance by over 35%. To keep with the theme of standouts, let’s focus our attention on it. Here it is below, to keep in mind as we proceed our analysis:

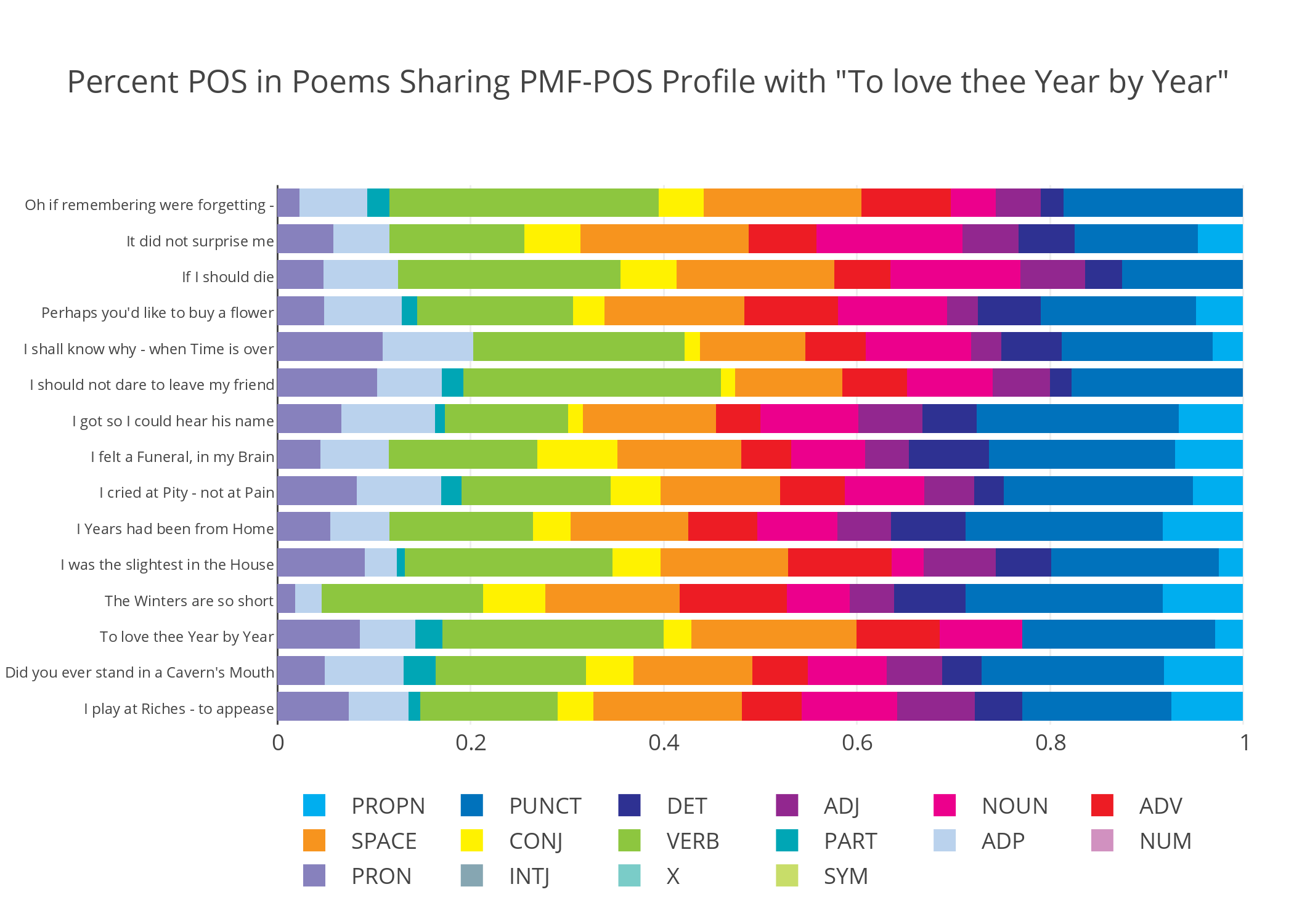
To love thee Year by Year—  
May less appear  
Than sacrifice, and cease—  
However, dear,  
Forever might be short, I thought to show—  
And so I pieced it, with a flower, now.

(“Emily Dickinson Archive”)

There is much more to be said about Fascicle 29, and it will be useful to keep in mind that at the end of this journey a return to Fascicle 29 – Dickinson’s provided form – will be appropriate, particularly to revise this model as needed. “To love thee Year by Year” matches POS profile 55 though, so let us turn our attention to it and other poems that match this POS profile. Below, we find that it is paired with fourteen other poems across thirteen separate fascicles. A list is provided for reference.

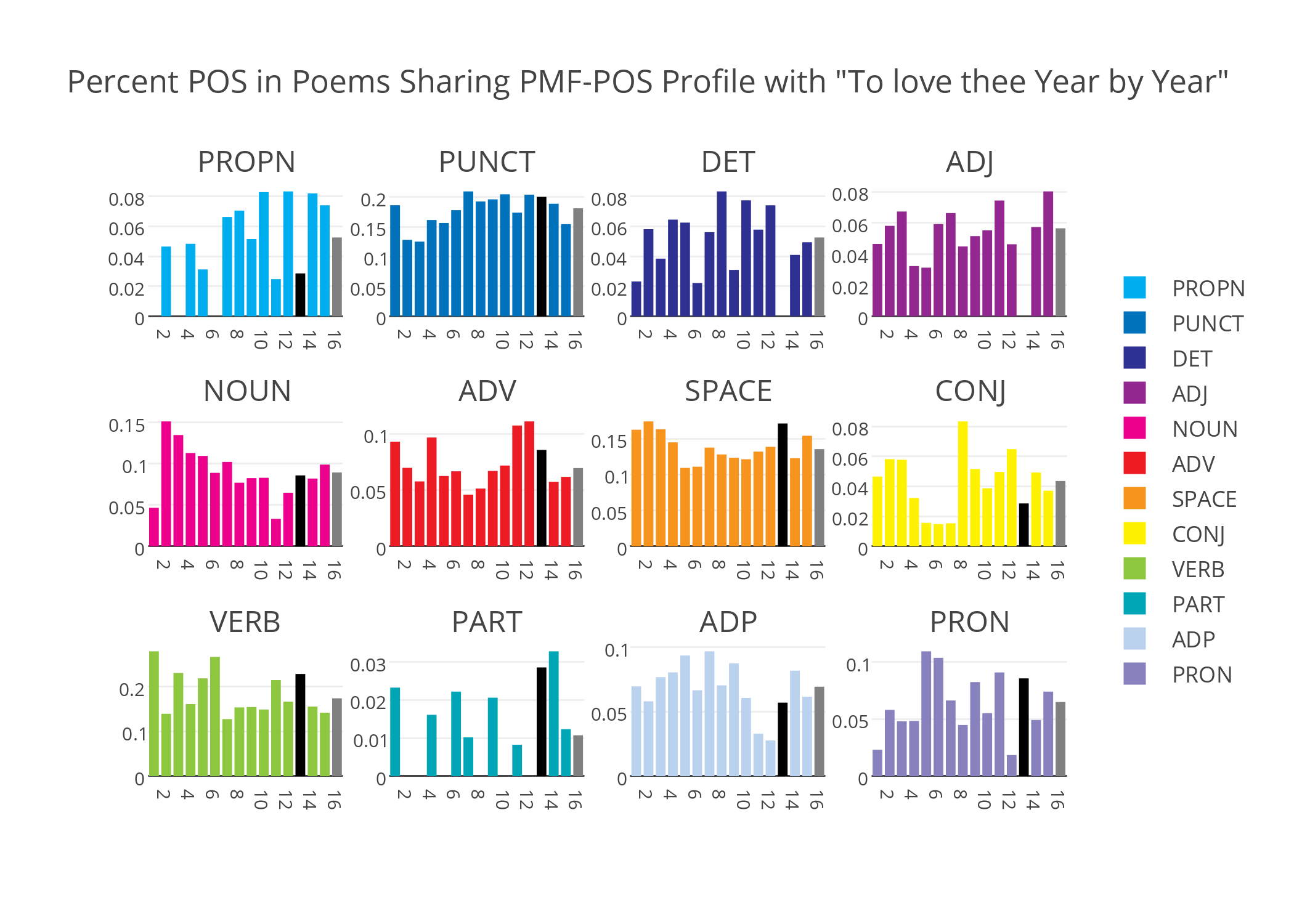


Now that we have focused on this new set of objects within the model, juxtaposed by POS tagging and PMF, we might wonder, “How are these poems related?” In fact, we might be tempted to just return to the act of reading to make this determination. But we are not quite ready for that. First, it will be helpful to more fully utilize this complex part-of-speech model in order to know where to look for the poems’ similarities and differences. Let’s take the familiar view of proportional part of speech usage of each poem and manipulate it in a few ways. First, here is the proportional view to get an initial perception of those POS dynamics.



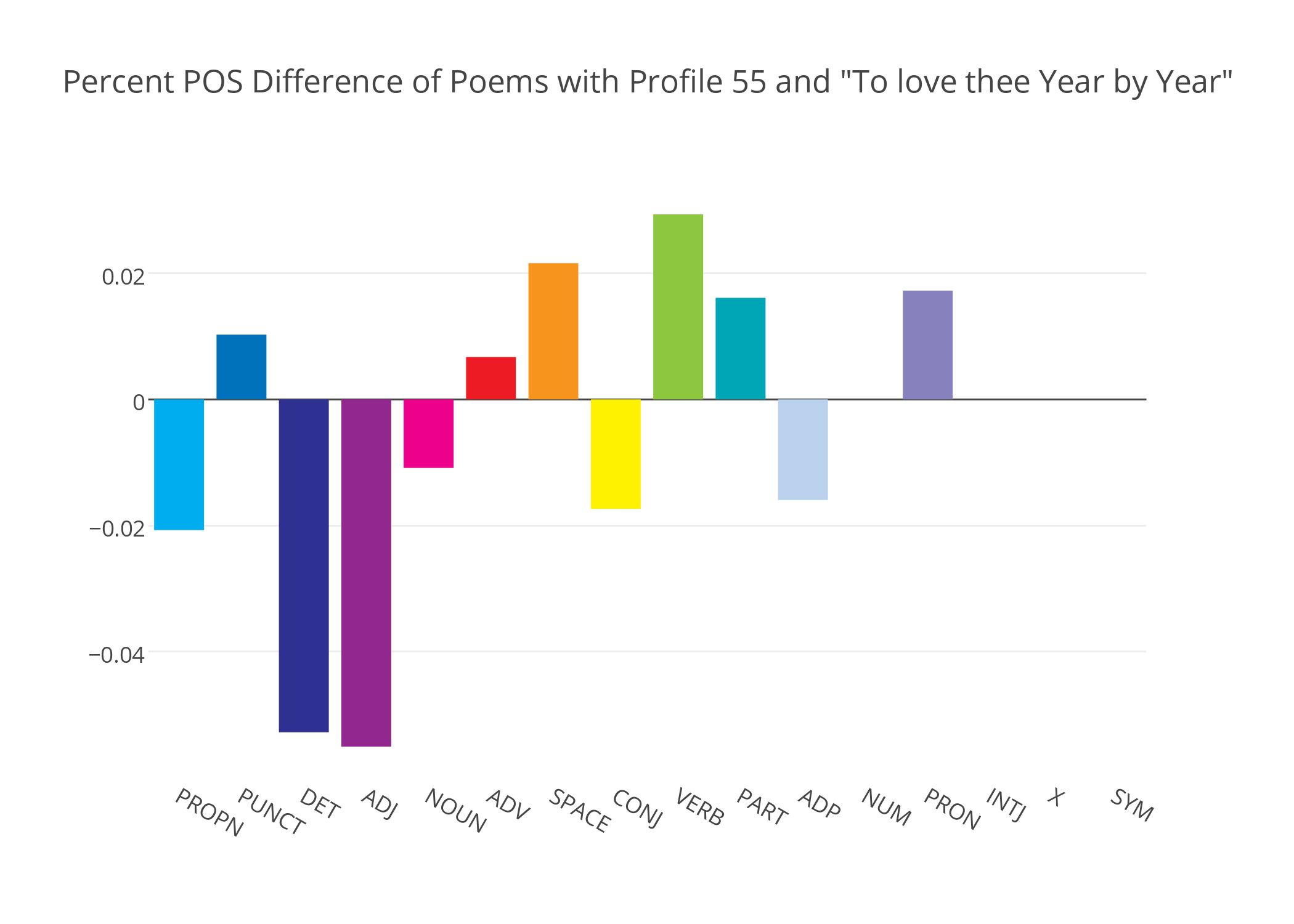
**Figure 8.** A proportional view of POS vectors of poems sharing PMF-POS profile 55

Next we divide these counts into their own separate graphs in order to get an exact sense of the POS dynamics moving through these poems in the manuscript-chronological order as established by Franklin and Johnson. “To love thee Year by Year” will be marked in black, and the PMF-POS mean profile vector will be marked in gray. The usual colors for POS are applied otherwise.



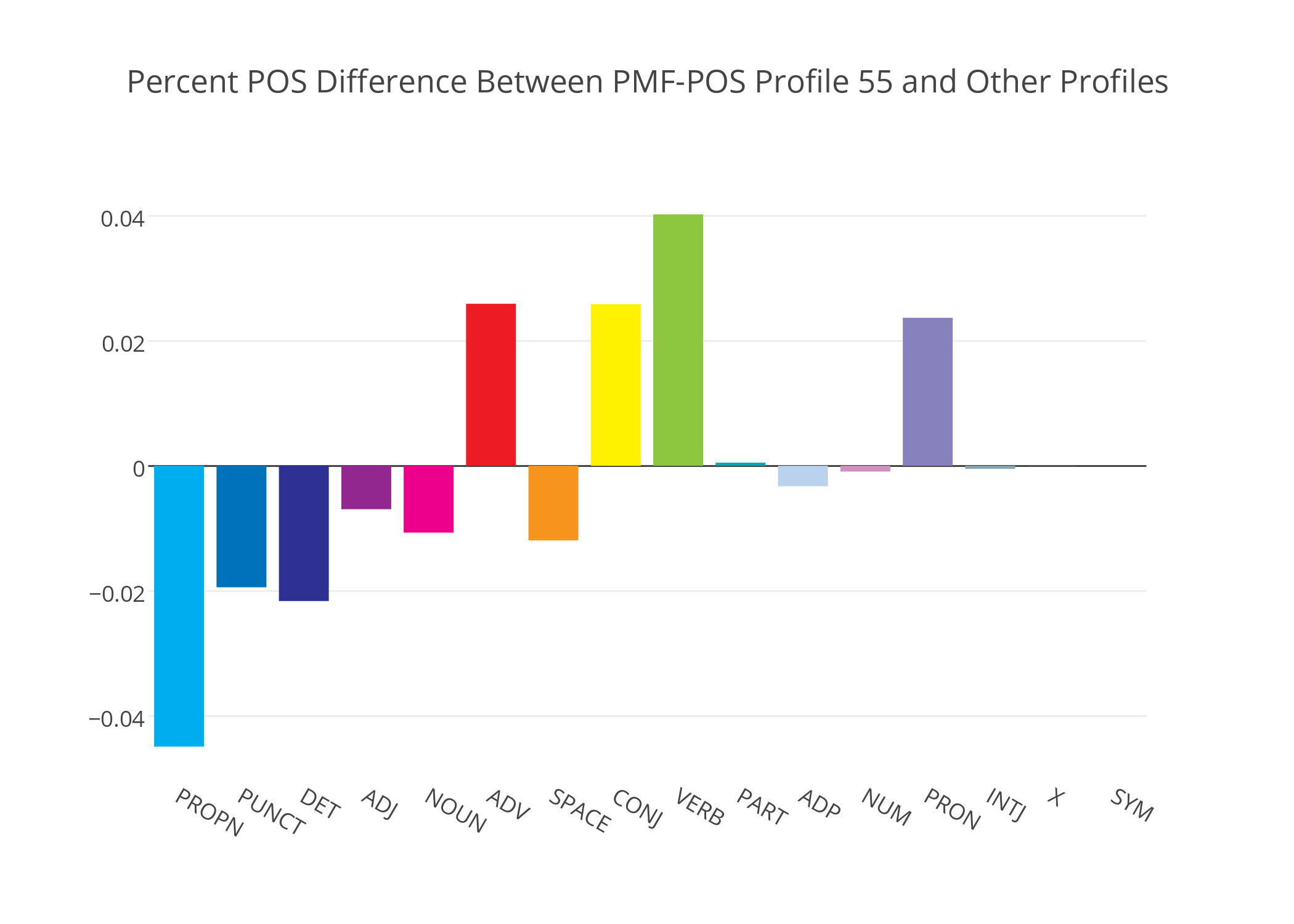
**Figure 9.** Proportional POS speech usage of all poems sharing PMF-POS profile 55, with “To love thee Year by Year” marked in black and PMF-POS mean profile vector #55 marked in gray. The x-axis denote poem numbers. This order can be recovered from Figure 10. POS categories not present are left out.

Now we can see precisely why and how “To love thee Year by Year” varies from its profile-mates. It is low in proper nouns, has little to no determinants or adjectives, and has relatively high adverbs, verbs, participles, and pronouns. Typical poem variance from the POS profile determined by the PMF algorithm notes that many of the poems exceed this profile in proper nouns, determinants, adjectives, nouns, adverbs and adpositions. The rest of the category variances are a bit mixed. You may note some key differences between these typical POS characteristics and “To love thee Year by Year”. Below they are visualized in a direct comparison that confirms those inferences.



**Figure 10.** A comparison between the POS vector of “To love thee Year by Year” and of other poems with PMF-POS profile 55

This visual indicates that “To love thee Year by Year” has lower amounts of proper nouns, determinants, adjectives, nouns, conjunctions, and adpositions than its profile-mates. We note that it features higher amounts of nouns, adverbs, spaces, verbs, participles, and pronouns. Verbs in this poem are notably higher than in the others. Up until now these POS count characteristics appeared to be random, but to return to a look at PMF-POS profile 55, we can put those characteristics into context.



**Figure 11.** A comparison between PMF-POS profile 55 and all other PMF identified POS profiles in Dickinson’s fascicle poems.

On average these poems all feature – relative to all fascicle poems – lower pronouns, punctuation, determinants, adjectives, nouns, and spaces and higher adverbs, conjunctions, verbs, and pronouns. Where these two visuals overlap the most is in the use of verbs. In fact, the fifteen poems of PMF-POS profile 55 feature 157 unique verbs. Perhaps there is some semantic relationship as well here? A proportional word cloud below represents raw frequency counts of the lemmatized verbs shared by these poems.[[8]](#footnote-8)



**Figure 12.** Word cloud of the top 25 lemmatized verbs shared by the poems of PMF-POS profile 55

Suggestive thematic pairs emerge in the top 25 of these verbs. They are listed here alongside their raw counts:

1. think (7), know (6), wish (3)
2. hear (5), look (4), hunt (2)
3. forget (5), remember (3)
4. die (4), cease (3), begin (2), live (2)
5. leave (3), come (3), depart (2), fly (2), dare (2)
6. bear (3), hold (3)
7. like (3)
8. say (3), make (2), deem (2), beat (2)

From these lists we get hints that these are indeed very active poems that contain discursive relationships on inner reflection, observation, memory and perhaps deal with thoughts on mortality. Returning finally to “To love thee Year by Year” we focus on the verb “cease” and realize that it is the fulcrum of the piece. It is the narrator’s attention to the end of a love or the end of a life that drives the poem. None of this is far flung from the well-studied themes of Dickinson’s poems, but remember that these are particular POS-induced aspects of such themes. In this case, that thematic is informed by relatively high usage of verbs. Comparing the high verb features of the first few poems (which are from the first ten fascicles) as seen in Figure 9 and the relatively higher verb usage of the first ten fascicles in Figure 1, this all starts to make a lot of sense. It also adds credibility to the matrix factorization model we have been exploring. At this point, we could return to examining these poems more closely looking at this PMF-POS grouping or we could return to Fascicle 29 itself. What this exploration has granted us though is an understanding of the latent factors that bind these poems together, as well as new contrasts to investigate. Never the less, this model of a model atop an archive is a new tool by which we might reconsider juxtapositions perhaps unnatural in normative archival contexts.

**Conclusion**

I began this writing on the archive with a curious phrase. I said, “Today, let’s go on a journey across an archive.” Just like Dinshaw’s asynchronous image, the contents of an archive, all carefully put into one container and described are considered in a timeframe of the archivist’s choosing. From the beginning of this writing we have been journeying past that static image in space and into another timeframe. I’ll call it “tomorrow” since our existence in physical space extends into the fourth dimension of time only by a metaphor of linear spatial order. Really, we have stepped outside of time. The use of higher dimensional data implicitly asks us to consider all timeframes and to do so outside of our comforting existence-related metaphors. To end, I repeat that invitation in seeming nonsensical form. “Tomorrow, let’s go on a journey across a model.” In a model, now is yesterday, today, and tomorrow. Improper juxtaposition and anachronism are no longer jarring. They are the rule. To be more concrete, we can see that a model acts as an archive in that it preserves the algorithm – the new, computational form – for accessing the materials it represents. And while this new topology of a model may be reconfigurable, it too has rules.

**Bibliography**

Armoza, Jonathan. “Part-of-Speech Profiling and Stylistic Textual Assessment Using Probabilistic Matrix Factorization,” 2016.

Dickinson, Emily. “Emily Dickinson Archive,” 2013. edickinson.org.

———. *The Manuscript Books of Emily Dickinson.* Edited by R.W. Franklin. Cambridge, MA: Harvard University Press, 1981.

Dinshaw, Carolyn. *How Soon Is Now?: Medieval Texts, Amateur Readers, and the Queerness of Time*. Duke University Press, 2012.

Halkidi, Maria, Yannis Batistakis, and Michalis Vazirgiannis. “On Clustering Validation Techniques.” *Journal of Intelligent Information Systems* 17, no. 2/3 (2001): 107–45.

Honnibal, Matthew. “Citation Information #272,” February 22, 2016. https://github.com/explosion/spaCy/issues/272.

———. *spaCy*, 2016. spacy.io.

Johnson, Thomas. *The Poems of Emily Dickinson*. The Belknap Press, 1955.

Koren, Yehuda, Robert Bell, and Chris Volinsky. “Matrix Factorization Techniques for Recommender Systems.” *IEEE Computer Society*, August 2009, 42–49.

Piper, Andrew. *Dreaming in Books: The Making of the Bibliographic Imagination in the Romantic Age*. The University of Chicago Press, 2009.

———. “Reading’s Refrain: From Bibliography to Topology.” *ELH* 80, no. 2 (Summer 2013).

Salakhutdinov, Ruslan, and Andriy Mnih. “Probabilistic Matrix Factorization.” *University of Toronto*, 2008.

Sinclair, Stéfan. *Voyant Tools* (version 2.0), 2016. http://voyant-tools.org.

Siskin, Clifford. *System: The Shaping of Modern Knowledge*. Kindle Edition. The MIT Press, 2016.

Zitnik, Marinka, and Blaz Zupan. “Nimfa: A Python Library for Nonnegative Matrix Factor- Ization.” *Journal of Machine Learning Research* 13 (2012): 849–53.

1. More info on the spaCy text parser can be found at <https://spacy.io/>. [↑](#footnote-ref-1)
2. Elements like punctuation, spaces, numbers, and symbols are considered parts of speech for spaCy’s purposes. [↑](#footnote-ref-2)
3. Inferring missing values has been found to produce the unfortunate result of making computation of these problems increasingly difficult. [↑](#footnote-ref-3)
4. Acceptable coefficients of these matrices are determined after a factorization has been calculated by the minimization of a related, error-based objective function. In the case of the method used in this writing, Probabilistic Matrix Factorization, it is the minimization of an error-based objective function that is informed by those prior statistical distributions representing the data and the two matrices resulting from factorization. The mathematics behind these methods is described in a separate paper, my “Part-of-Speech Profiling and Stylistic Textual Assessment Using Probabilistic Matrix Factorization” (listed in the bibliography and available upon request) and also in Salakhutdinov and Mnih’s “Probabilistic Matrix Factorization.” [↑](#footnote-ref-4)
5. This is the metaphor employed by a set of researchers using the matrix factorization code library, “Nimfa,” used for the PMF results below. See bibliography and <http://nimfa.biolab.si/>. [↑](#footnote-ref-5)
6. The prior statistical distributions given to the method as assumptions for this modeling demonstration are nothing but standard bell-curves (a.k.a. normal or Gaussian distributions). Remember, priors offer information to a modeling method of how it should expect incoming data to roughly be distributed. This expectation is also quickly overtaken as more data is consumed. [↑](#footnote-ref-6)
7. The distances used in these comparisons are known as “Jensen-Shannon” distances, a calculation that provides a scalar value as a distance between two vectors. [↑](#footnote-ref-7)
8. Cirrus cloud visual provided by “Voyant Tools” at <http://voyant-tools.org/>. [↑](#footnote-ref-8)