DHQ: Digital Humanities Quarterly
2016
Volume 10 Number 4

Language DNA: Visualizing a Language Decomposition

Adam James Bradley <adam_dot_bradley_at_uwaterloo_dot_ca>, University of Waterloo
Travis Kirton <tkirton_at_gmail_dot_com>, University of Calgary
Mark Hancock <mark_dot_hancock_at_uwaterloo_dot_ca>, University of Waterloo
Sheelagh Carpendale <sheelagh_at_ucalgary_dot_ca>, University of Calgary

Abstract

In the Digital Humanities, there is a fast-growing body of research that uses data visualization to explore the structures of language. While new techniques are proliferating they still fall short of offering whole language experimentation. We provide a mathematical technique that maps words and symbols to ordered unique numerical values, showing that this mapping is one-to-one and onto. We demonstrate this technique through linear, planar, and volumetric visualizations of data sets as large as the Oxford English Dictionary and as small as a single poem. The visualizations of this space have been designed to engage the viewer in the analogic practice of comparison already in use by literary critics but on a scale inaccessible by other means. We studied our visualization with expert participants from many fields including English studies, Information Visualization, Human-Computer Interaction, and Computer Graphics. We present our findings from this study and discuss both the criticisms and validations of our approach.

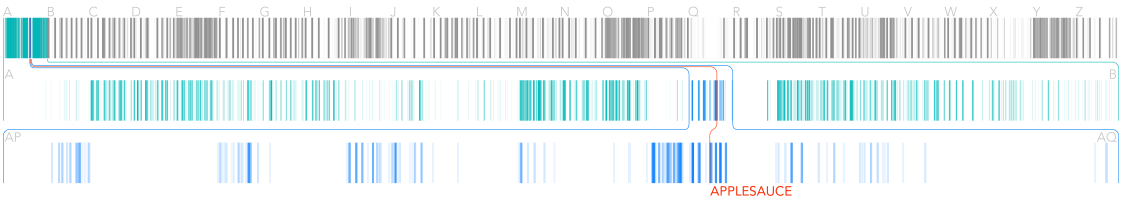

Figure 1. The English language in 1D along the top row. Showing zoom levels to find the word “applesauce”

1 INTRODUCTION

One of the goals of the literary critic is to analyze language and its embedded complexity. For example, when literary critics examine a poem’s form, they consider many characteristics of the words it contains, including the similarities and differences in orthography, sound, the visible pattern it produces, rhythmic structure, and countless others. Nearly all of this information is available through visual inspection of the poem and contained in what may already be our greatest visualization technique — the written word. However, this same inspection carries with it the biases introduced by the semantic meanings of the words themselves: it is difficult to pay attention to the structural parts of the word “apple” without imagining the fruit it represents. 1

One possible method of aiding in the process of literary criticism is to provide an alternate representation [Simon 1988] of the words contained within the text to help separate the meaning from the structure, or to provide a different vantage point with which to examine the work. By providing this alternate view, structural aspects of the poem that may be difficult to recognize when obfuscated by the meaning of the words themselves may be easier to analyze and observe. We caution that this understanding will likely not be valuable unless a connection can be bridged between the alternate representation and the original piece of literature, and we stress that this approach is to be considered an augmentation 2

to existing practices. The goal of the research presented in this paper is to provide an information visualization that provides alternate representations of literature that will allow the critic to recover the original work for analogic purposes.

There have been a variety of examples in the digital humanities and information visualization disciplines that provide alternate representations of language, such as Word Clouds [Viegas 2008] and Word Trees [Wattenberg 2008], but the authors are not aware of any examples that have as a priority the ability to discern the text itself from the visualization. Instead, these examples provide a means of examining text with prior knowledge of the contents therein. Nonetheless, it is an open question whether this change in representation would prove useful to the trained literary critic. 3

In this paper, we present a design study of an information visualization that is a recoverable representation of language. Specifically, we present the design of a visualization we have named Language DNA (L-DNA) that visually encodes any symbol system, in our case the letters and phonemes of words in the English language, and a qualitative evaluation with several literary critics and designers to investigate the need and use of such a visualization. 4

2 RELATED WORK

Our study is an examination of how to visualize language in ways that can build ontologies of words based on the needs of the literary critic. Much work has been done in the area of text visualization but none have yet approached a level that can produce whole language interactions. Our work builds on previous text visualization techniques while expanding on the scope of information that can be shown in our system, Language DNA. 5

2.1 Text Visualization

Many previous text visualizations focused on some form of text. For instance there are visualizations of documents [Collins 2009a] [Paley 2002], of selected subsets of words [DeCamp 2005] [Hetzler 1998], and repetition in context [Hearst 1995] [van Ham 2009] [Wattenberg 2008]. There are also visualizations for grouping and clustering documents [Collins 2009b] [Wise 1995]. However, they do not approach whole language visualization. Most approaches to text visualization whether coming from the Information Visualization (InfoVis) community or the digital humanities, are often aesthetically pleasing, but have not been demonstrated to have utility for literary criticism or linguistics. This limitation can be seen in text visualization projects such as tag clouds [Lee 2010] [Viegas 2008] [Viegas 2009] word associations [Yatani 2011] [van Ham 2009], topic modeling [Chen 2006] [Gardiner 2010] [Gretarsson 2011] [Hall 2008], and text document comparison [Collins 2009a] [Lin 1992] [Smalheiser 2009] [Wise 1995]. There are text visualization projects that fall into the category of linguistics that have had some success. [Rohrdantz 2012] provide a visual analysis of suffixes in English, and the Google N-Gram[1] viewer has given access to historical usage data previously unattainable. 6

2.2 Visualizing Language in the Digital Humanities

Within the domain of digital humanities, there are three main streams of visualization that tend to pervade the literature. All three types are based in text analysis but approach the problem from different directions. The first is the GIS type tools for organizing spatial data in the humanities [Jessop 2006] [Jessop 2008] [Wood 1992], specific examples of this type of work include Bingenheimer et al.'s visualization of the biographies of Buddhist monks [Bingenheimer 2009], as well as Valley of the Shadow [Valley 2015], in which there are examples of multiple types of quantitative and qualitative data presented in a GIS format. 7

The second category involves tools used to augment reading, text analysis tools that highlight relationships in texts, often with visualizations. This category includes work such as Clement's distant reading of Gertrude Stein [Clement 2008], and Wattenberg et al.'s work on using tree visualizations for visual concordances [Wattenberg 2008]. Projects such as TAPoR [Tapor 2015] also provide data mining tools for literary and textual analyses. The aim of these projects is to allow for interesting portions of large texts to be marked up and visualized for comparison. 8

The third category of research is the straight visualization projects which are of two types: The first are the projects that use toolkits such as Voyant tools [Voyant 2015] to accomplish their research task and the second are the projects that create bespoke visualizations specific to their research problem. The second type of these includes Moretti's work on 9

visualizing interactions and character movements in literature [Moretti 2005], Bingerheimer et al.'s social network visualization from TEI data, which uses text encoded data to produce networked relationships among textual elements [Bingenheimer 2006], and Gould's use of network tools for historical data analysis [Gould 2003]. This can also be seen in the TextArc project, which uses radial graphs to show contents and relationships between words in texts [Paley 2002], and Writing without Words [Posavec 2015], Posavec's beautiful graphs of sentence structures and themes in Kerouac's "On the Road". Perhaps the most exciting and cognate study is Alexander's work on the Oxford Historical Thesaurus and mapping sonic relationships in language over time [Alexander 2012]. Also, we find a movement to projects that use art to approach these problems, highlighting that aesthetics may be itself a critical tool. Installations such as TextRain [TextRain 2015], an interactive art piece where gallery goers interact with falling text, and Word Collider [WordCollider 2015], an artistic project modeled after pictures from CERN's Large Hadron Collider, where parts of words when "smashed" together produce effects like those seen in images from particle accelerators.

The development of our technique was motivated by the fact that existing visualizations and text analysis tools, while usually aesthetic pleasing, cannot easily be used for the types of analyses expected or desirable for literary criticism or linguistics. Our Language DNA is an attempt to build a system that can handle multiple levels of information to display complex structural and content-related relationships within texts. 10

3 MOTIVATION — A LITERARY CRITIC'S PERSPECTIVE

In his book *Anatomy of Criticism*, Northrop Frye asks the question: "what if criticism is a science as well as an art?" [Frye 2000] The development of the visualization technique presented here was driven by this exact question. Our intention is to facilitate a different approach to Frye's domain-specific question about literary criticism: can we use mathematical and visualization techniques to incorporate science into a literary criticism? These questions have a long history in the humanities and were born out of the general notion that arose in the renaissance when the likes of Locke, Bacon, and Descartes defined a practice that separated and solidified science from the humanities. The 20th century structuralists suggest that this position is reconcilable, that all literature and language is systematic. This idea — that language is systematic — is approachable via visualization. 11

The visualization algorithm we present can accommodate as much or as little information that a critic could want, giving the possibility of visualizing as little as a single letter, as much as entire corpora. If we are to imagine a space where the literary critic or linguist can experiment using a language-based mathematics, it must have these three characteristics: 12

3.1 Consistency and Reversibility

First, a visualization algorithm is needed to encode language such that we can create a space that is both consistent and reversible. In mathematical terms, this would be referred to as one-to-one and onto. The need for consistency and reversibility arises from the requirement in the analysis process of preserving the ability for human interpretation of the words. In order for the visualization to be "readable" by a critic, each word must be consistently mapped and that mapping must be reversible into the original work. Most existing visualization techniques distort the original texts without providing an avenue for reconstituting them. This is problematic when studying things like poetics where the spatial component of the text is integral to its meaning. 13

3.2 Infinite Plotting Space

Second, it is important that the algorithm uses a plotting space that is infinite. Imagine that we were to approach the problem of metaphor. As an example, in theory we do not know if the chain of meaning created by metaphoric relations between words is finite. It stands to reason then, that without an understanding of the full requirements of a system that an infinite space is a safe decision. If we are to start to approach these types of questions our lack of knowledge should not limit the size of possibilities. Specifically, the critic must be able to analyze words and literature that are perhaps not known to the visualization designer. With our technique, we can encode anywhere from one letter to an entire language, to the entire literatures of one language, and even multiple languages, in a space where each point belongs to one individual piece of original information. The infinite space means that in creating "experiments" the literary critic is not limited to our present understanding of language. 14

3.3 Layering

The third requirement is the need for the ability to layer symbols in order to make comparisons. For instance, it should be possible to overlay a poem within the context of the entire language or other poetry. The need for layering arises from the analogic basis of most types of literary critical and linguistic inquiries. This should extend to any types of symbols, as comparisons are not always rooted in the Latin alphabet. Based on what we recognize as the possible requirements of such a system, we designed our Language DNA visualization with the three characteristics of consistency and reversibility, infinite plotting space, and layering in mind. 15

4 DEFINING LANGUAGE DNA

An important property of our mapping is that the words be recoverable from the visual space. Mathematically, this requires that the mapping be a bijection (i.e., that words both map to a unique place in the visual space, and that each point in the visual space maps back to a word). As an example of the technique we introduce a mathematical translation of words to numbers that relies on the lexicographical ordering of letters. This is essentially a mapping of alphabetical order and is one of possibly infinite ways to group the data. We have chosen this technique as a first demonstration because we are all familiar with the way we order a dictionary, but we must stress that we can map the data many different ways. We define the mapping g so that each letter is mapped to its position in the alphabet, as follows: 16

$$g : A \rightarrow \mathbb{Z} \quad (17)$$

where A is the set of alphabetical characters $\{a, b \ldots, z\}$ and: 18

$$g(a) = 1,\ g(b) = 2,\ \ldots,\ g(z) = 26 \quad (19)$$

Note that this mapping is currently written using base 10 numbers for the integers (1 to 26, with an implied 0 for no character), but our mapping requires a base 27 representation (or more generally base N+1, where N is the number of characters in the language), which for convenience we will symbolically represent as follows: 20

$1_{10} = a_{27}$ (i.e., 1 in base 10 is represented as ‘a’ in base 27) 21
$2_{10} = b_{27}$
...
$26_{10} = z_{27}$

Thus, we can define our mapping of words to a one-dimensional number line as follows: 22

$$f: W \rightarrow (0,1) \quad (23)$$

where W is the set of alphabetical words (e.g.,“apple”, “dog”, “the”, etc.) such that for each $w \in W$, $w = x_1\ x_2 \ldots x_n$, and: 24

$$f(w) = 0.(x_1)g(x_2) \ldots g(x_n) \quad (25)$$

For example, for w = “dog”, we have x_1 = ‘d’, x_2 = ‘o’, x_3 = ‘g’, therefore: 26

$$f(\text{"dog"}) = 0.dog \quad (27)$$

Note that this is a base 27 number, but could be converted to base 10: 28

$$0.dog_{27} = 4\text{x}27^{-1} + 15\text{x}27^{-2} + 7\text{x}27^{-3} = 0.1690799_{10} \quad (29)$$

Figure 2. All words in the English language visualized along a number line from 0 to 1, and the specific word "dog" (represented as a vertical line in red) would appear 16.9% of the way in.

If we relax the restriction that each word needs to end (i.e., we allow words to have an infinite sequence of letters), it becomes clear that f is a bijection, since every word generates a unique base 27 representation (one-to-one: the property that if two words map to the same number, they must be the same word) and each number between 0 and 1 can be converted to base 27 to recover the sequence of letters (onto: the property that every number has a word that can map onto it). 30

4.2 Using the 2D Visual Space

The mapping above describes how an arbitrary word can be mapped onto a number line, which already allows the visual mapping of words onto an axis in 1D space (similar to a lexicographical axis). Here, we describe a method, inspired by Cantor's Diagonalization to map an individual word onto 2D space directly (and more generally onto n-dimensional space). 31

We can split the word in two by considering every other character, for instance *InFoViS* would become *IFVS* and *noi*. Thus, we can take the base 27 representation of the mapped word and create two dimensions as follows: 32

$$f_2: W \rightarrow (0,1)^2 \quad 33$$

$$f_2(w) = f_2(x_1 x_2 \ldots x_{2n}) = (f(x_1 x_3 \ldots x_{2n-1}), (f(x_2 x_4 \ldots x_{2n}))$$

For example, if our word is "applesauce" (Fig. 3): 34

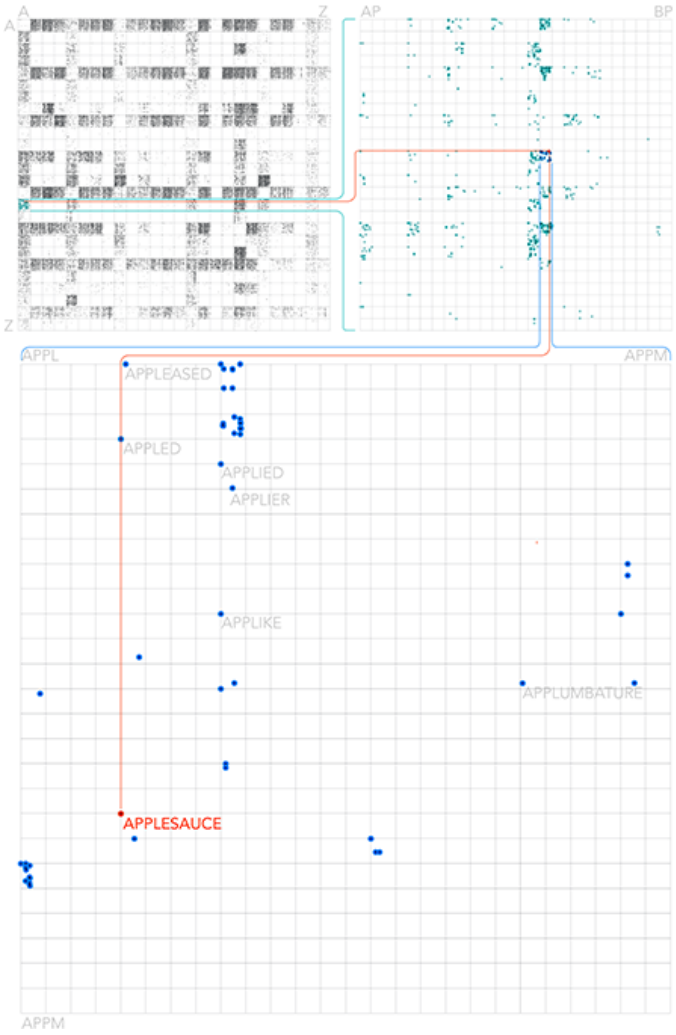

Figure 3. The entire English language can also be visualized in 2D (top-left). The result is what appears as a grid of letter pairs, so that one can zoom into words that begin with AP (top-right), and within that square, words that begin with APPL (bottom).

f_2 ("applesauce") = (0.$apeac$, 0.$plsue$) 35

which, in base 10 would be: 36

f_2 ("applesauce") = (0.0592410,0.6100587) 37

This mapping can easily be extended to n-dimensions by taking every n^{th} character in the base 27 representation of (ω). f_2 is also clearly a bijection, because every word can be split into alternating characters to generate two base 27 representations (one-to-one) and each pair of numbers between 0 and 1 can be converted to base 27 to recover the two parts of the word, which can then be reassembled (onto). Thus, every word in the English language can be mapped onto a 2D space using f_2, and every 2D point can be mapped to a “word”, where a word is a sequence of possibly infinite letters which may well not have associated semantics. Note that this mapping does not account for things like homonyms, but with a simple addition to the mapping we could easily differentiate words for any number of their ontological characteristics. 38

4.3 A Note on Scale

Since whole natural languages are immense, it is important to discuss scale, both of what is being visualized and the size of the resulting visualization. We can base a visualization size calculation on the number of words being visualized, and then determine the length of a 1D L-DNA visualization that draws at a density of a single pixel for each word or unit 39

(it is important to remember that these calculations are for orthography, they will change depending on the symbol system used). Two measures are needed to accomplish this: the smallest and the largest distance between two words. Since our algorithm already normalizes words in 1D to be between 0 and 1, we can assume that the difference between the largest and smallest words is approximately 1.0 (with the words 'a' and 'zygote', this is already correct to 1 decimal place). In our analysis of words from the Oxford English Dictionary (OED), the two closest words using our algorithm are "abandoner" and "abandoning", with the first seven letters in common and the next being very close in the alphabet. The difference in values from our algorithm for these two words is:

$$0.abandoning_{27} - 0.abandoner_{27} = 1.37 \times 10^{-11}{}_{10} \quad (40)$$

Thus, to present a number line from 0 to 1 with numbers only 1.37×10^{-11} apart represented as different pixels would require: (41)

$$1.0 \div (1.37 \times 10^{-11}) = 73.1 \text{ billion pixels} \quad (42)$$

Note that in 2D, our algorithm fairs far better. This same pair of words would be broken down into two pairs of coordinates: (43)

$$(0.aadnn_{27}, 0.bnoig_{27}) \text{ and } (0.aadnr_{27}, 0.bnoe_{27}) \quad (44)$$

which has at most four letters in common for each dimension and would require only: (45)

$$1.0 \div (0.aadnr_{27} - 0.aadnn_{27}) = 1.0 \div (2.79\text{x}10^{-7}) = 3.6 \text{ million pixels} \quad (46)$$

To put this into perspective, a 1D visualization using our algorithm would require the width of 38.1 million 1080p screens (1920 × 1080 pixels) placed side-by-side, and a 2D visualization would require 6.2 million 1080p screens arranged to form a rectangle. (47)

5 ILLUSTRATING L-DNA

We start with three examples to demonstrate how L-DNA can be used to visualize language. The first is a dictionary mapping for the English language, the second is a view of multiple languages, and the third is a mapping of English phonemes to illustrate the applicability of this approach to any set of symbols[2]. (48)

5.1 Visualizing the Oxford English Dictionary

Fig. 4 shows all 370,624 words parsed with criteria that eliminate diacritics and punctuation the Oxford English Dictionary[3] rendered using the algorithm described above (i.e., using the coordinates provided by f_2. The result is a mapping that privileges the first two letters of each word. That is, the x-axis can be read as an alphabetical ordering of the first letters of words, and the y-axis can be read as an alphabetical ordering of the second letters. This makes the top left box "AA", where you would find words such as "Aardvark" (note that few words in English begin with two A's, which is why this box is quite sparse). This property is recursive, so that within each box, the third and fourth letters are similarly privileged. For example, in the "BA" box, there is an "NA" box, which has another "NA" box that contains the word "BANANA". This initial visualization shows how we can start to understand where each word belongs within the 2d whole. (49)

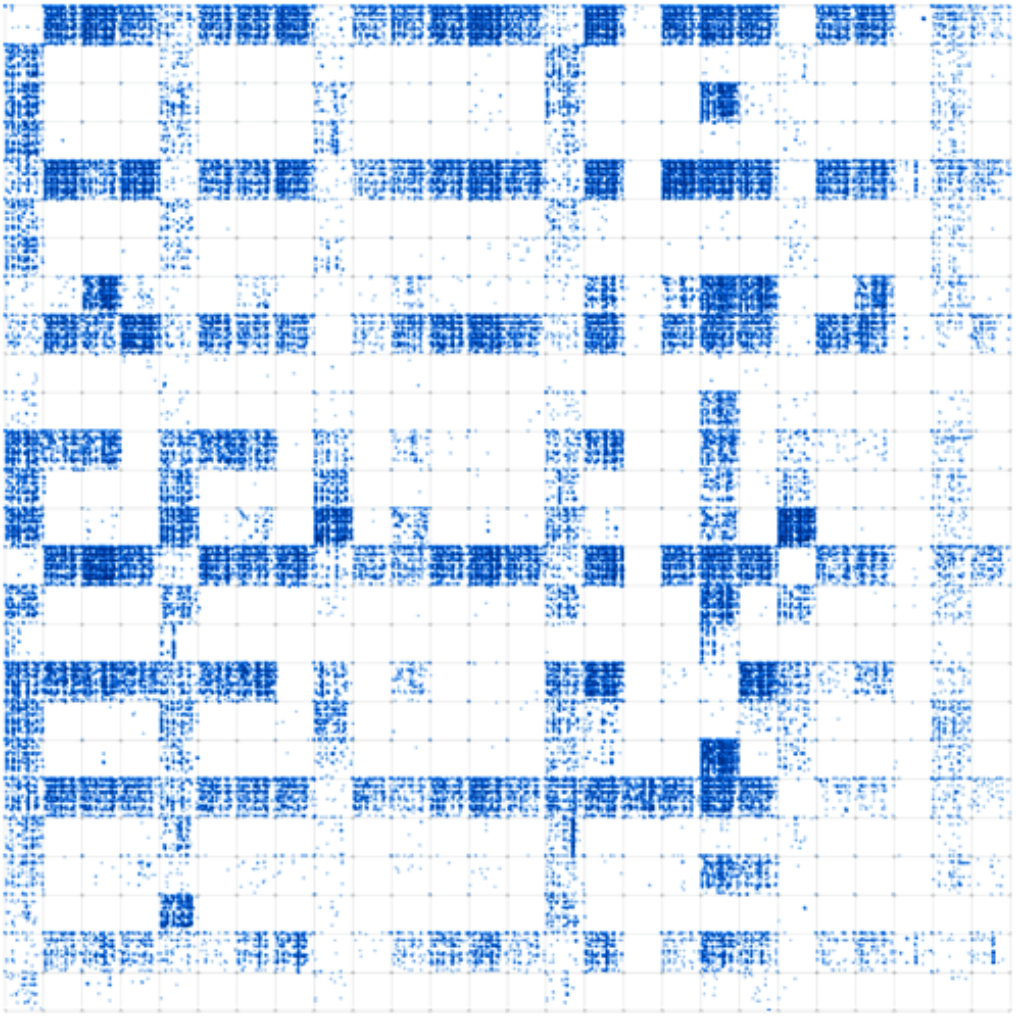

Figure 4. The Oxford English Dictionary (OED2) in 2D L-DNA

Because our algorithm privileges the spelling of words, this 2D representation can be thought of as a form of 2D orthography (specifically spelling rules). It is essentially a two-dimensional layout of alphabetical order. This version of L-DNA reveals a “bird’s-eye view” of the language that was not previously available to the literary critic, linguist, or lexicographer; a critic could previously flip through a dictionary’s pages or even a list of ordered English words, but this visualization instead provides a new 2d spatial location for each word in this dictionary. 50

5.2 The Multiple Language L-DNA View

Our second example compares multiple languages (English, French, German, and Spanish). Fig. 5 shows these four languages each represented in 1D on the 0 to 1 number line using our algorithm, stacked for comparison. Visual inspection reveals a similar sparseness in the ‘Q’ portion of the line for all languages (i.e., few words in any of these languages begin with ‘Q’ and any letter other than ‘U’), but additional sparseness in French, German, and Spanish exists near the end of the alphabet (‘W’, ‘X’, ‘Y’). 51

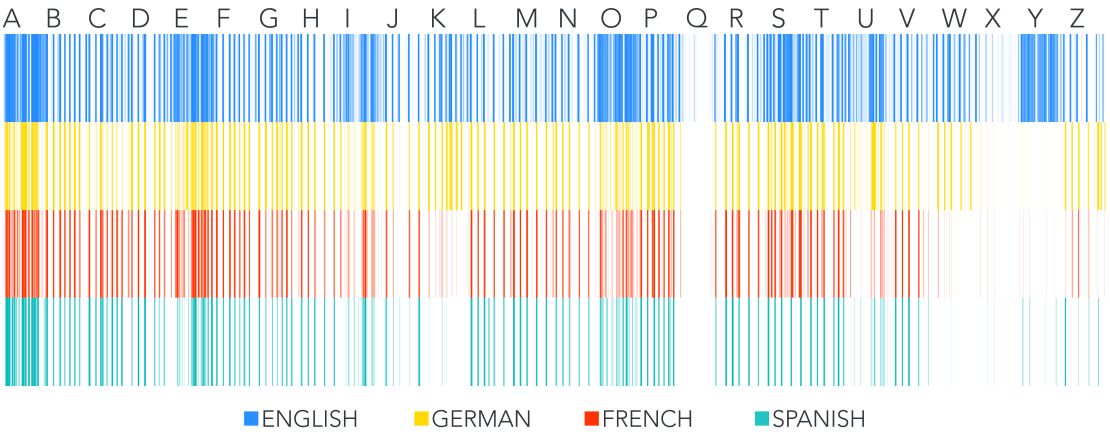

Figure 5. Four languages in 1D stacked.

Fig. 6 also shows a side-by-side comparison of multiple languages in 2D, and Fig. 7 overlays these four languages. Fig. 8 shows a close-up of the AL region of the overlaid image. These side-by-side comparisons or overlays allow for elementary analogic comparisons and can be expanded on with more complex symbol encoding.

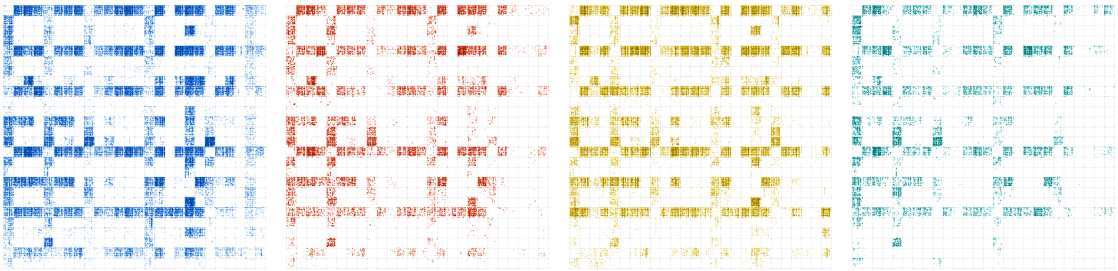

Figure 6. 2D L-DNA of English (blue), French (red), German (amber), Spanish (aqua)

Figure 7. Overlaying the dictionaries of the 4 languages (English, French, German and Spanish) in 2D L-DNA.

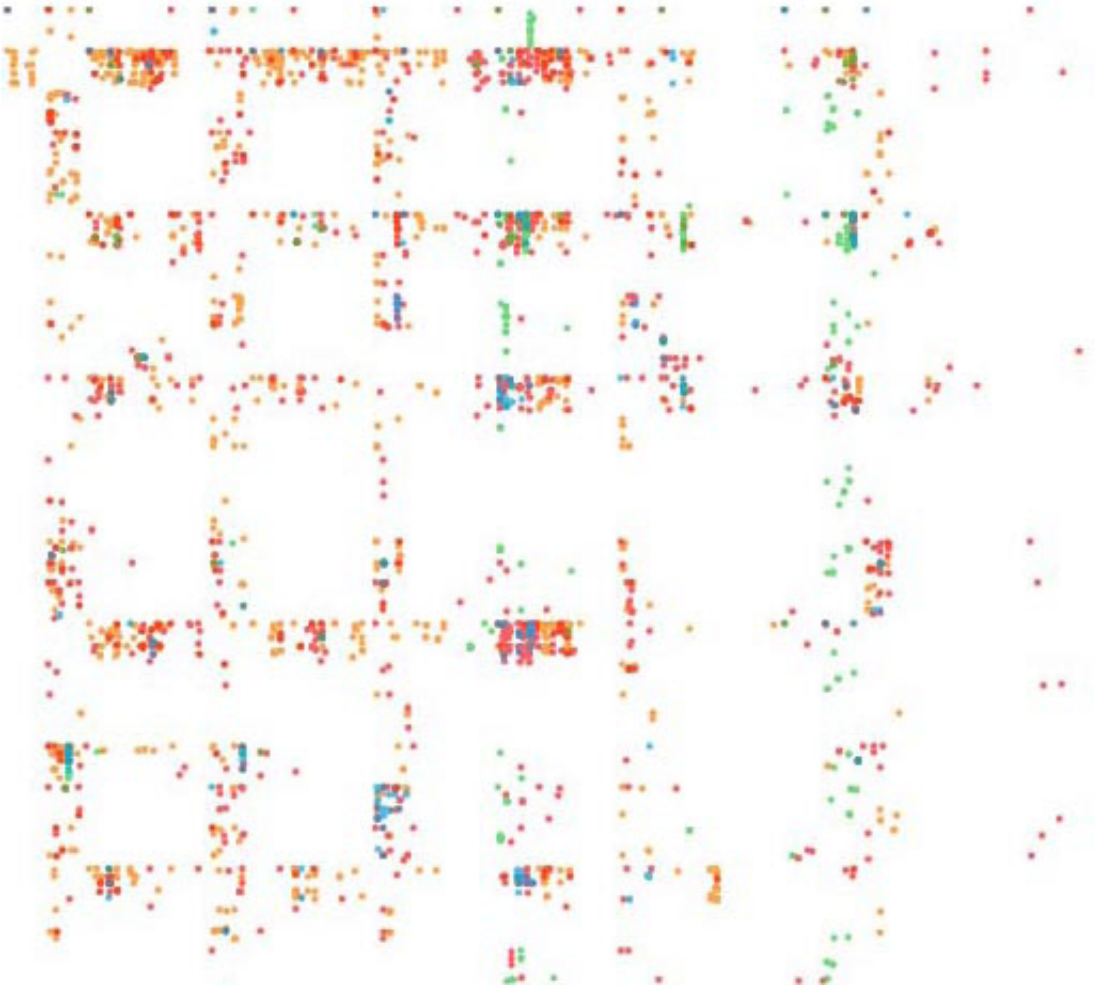

Figure 8. Close up of the subsection AL from Fig. 7

The above images were generated with the constraint that we only had access to open source dictionaries [Dict 2015] for languages other than English (for which we have university-wide access to the OED). The French (red) has 197,954 words, the English (blue) has 370,624 words, the German (yellow) 425,501 words, and the Spanish (green) has 160,442 words. 53

5.3 English Phonemes in L-DNA

We chose the next example (Fig. 9), English phonemes, to demonstrate the robustness of the technique to arbitrary symbolic representations of language, and to create an analogue between the spellings of words and the sounds of words. The mapping is organized in like sound units: vowels (e.g., “AA, AE”), semivowels (e.g., “W,Y”), stops (e.g., “B,D,K”), affricates (e.g., “C, H, JH”), fricatives (e.g., “D, SH, V”), aspirates (e.g., “HH”), liquids (e.g., “L, R”), and nasals (e.g., “M, N, NG”). 54

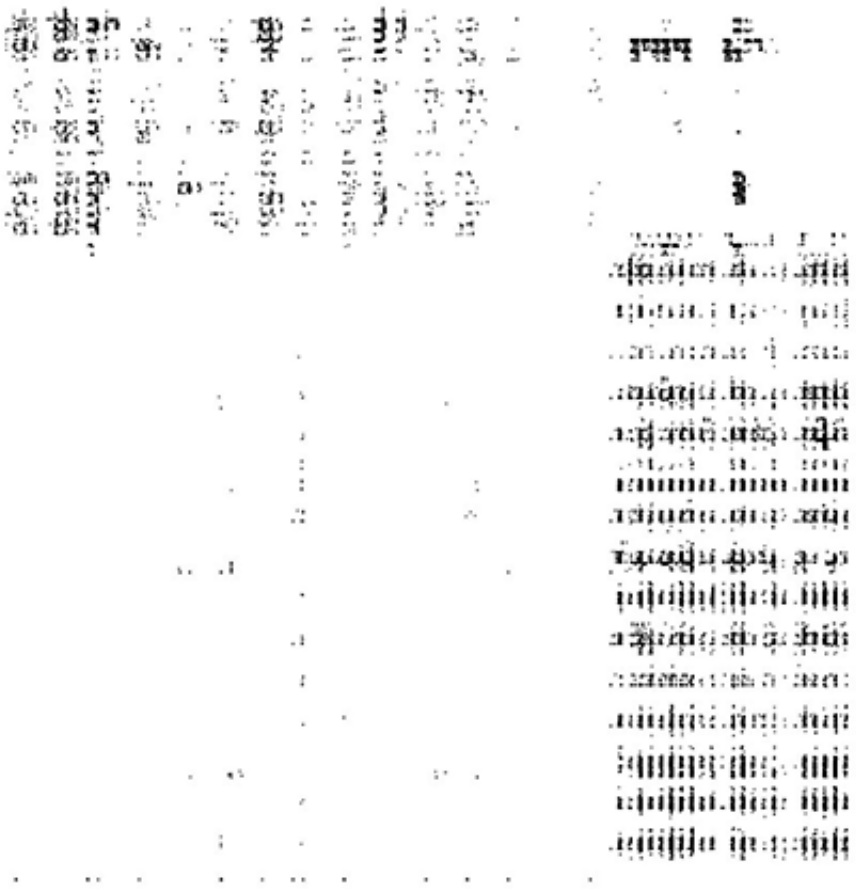

Figure 9. English Phonemes showing the Phoneme map of 370,624 words used from the OED2 in 2D mapped by sound types (i.e. fricatives, stops, etc.)

By organizing words into phonemes, some interesting observations can be made. It appears clear that a portion of the phonemes are used primarily for the first syllable and another distinct set is used primarily for the second syllable. This can be observed through the densely populated top and right columns, with the majority of the bottom-left part of the image containing almost no English words. In addition, the top-right corner is mostly empty, with the exception of a few very dense groups, representing the few phonemes that are used for both the first and second syllables. 55

5.4 Poetry

The final example that we created was to insert a single poem into the space that we created for English words and phonemes. This is a first step in being able to use these spaces for analogic comparisons. 56

Some interesting patterns can be observed in the poem through visualizing it in this manner. Firstly, in terms of orthography (Fig. 10) it becomes possible to identify visual rhymes by cluster groups within the image. In Fig. 11, this same phoneme visualization can be used to identify rhyming patterns within a poem. As the phonemes group together it is possible to see the types of sounds being repeated in the piece. Although this is easy to do with a 16 line sonnet, it becomes much more difficult with a poem of any significant length (e.g. Milton's 10,000 + lines of verse in "Paradise Lost") and this technique could help to highlight "macro-structures" in poetry. Each diagram below is laid out in two dimensions. This is a decision made during the encoding process and can be n-dimensional based on the amount of information you wish to build into the model. In this case we have chosen to show 2d representations for simplicity. In Figure 10 we present what we label our alphabetical order visualization where we represent words by their spellings. A visualization of spelling alone may not lead to many insights, but is useful for simple demonstrations. One area where this simple encoding could be used would be to visually compare irregular spellings in Elizabethan drama. It would provide quick visual access to the places in texts that differed and needed an editor's attention. The real power in this 57

method comes from being able to encode as many connections as desired. Work has been done in the digital humanities and computer science in the last few years in word embedding models and the consistency of our method could aid in the process of detecting connections in texts by using vectors. In figure 11, we have graphed phonemes in two dimensions. Any highlights that form vertical lines are showing alliteration in the poem. An example is in the line from Donne's poem: "Or like a thiefe, which till deaths doome be read". The words "deaths" and "doom" line up vertically to indicate alliteration in this particular encoding. If we wanted to visualize rhymes we would simply encode the phonemes in reverse, privileging the final phoneme and we would generate a similar graph. The n-dimensional nature of the models allows for as much or as little data coding as needed, including relations between words. The only limitations on the questions that can be asked are the imagination of the analyst.

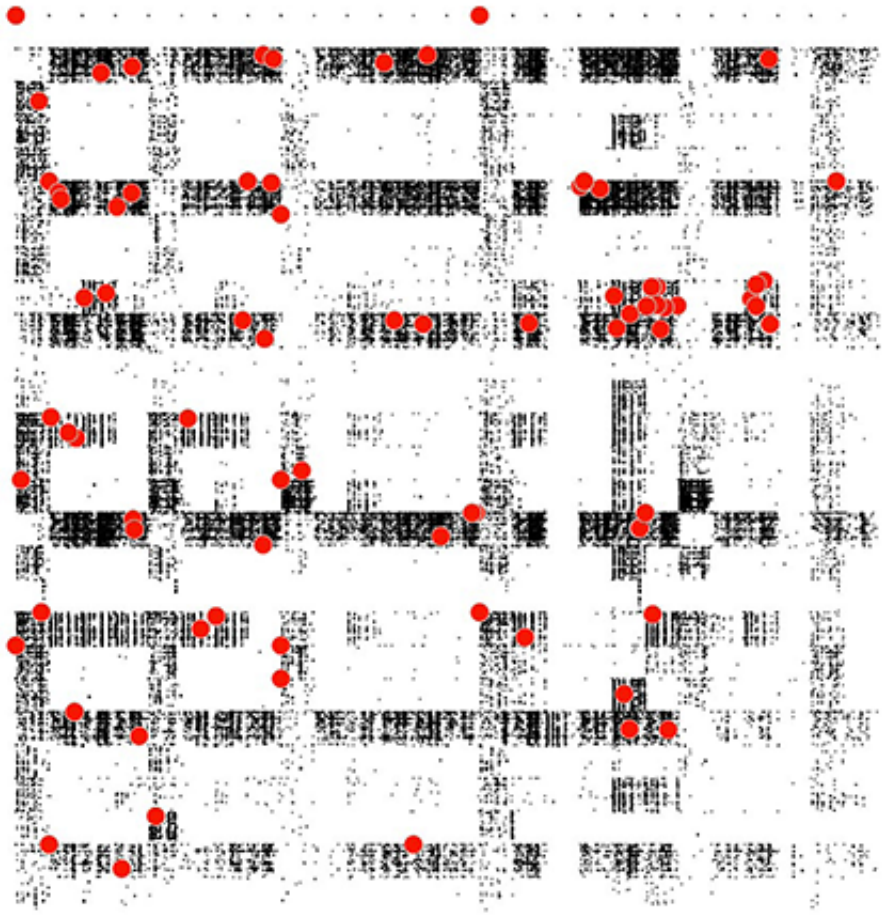

Figure 10. The text of John Donne's "O my black soul" sonnet visualized with the OED.

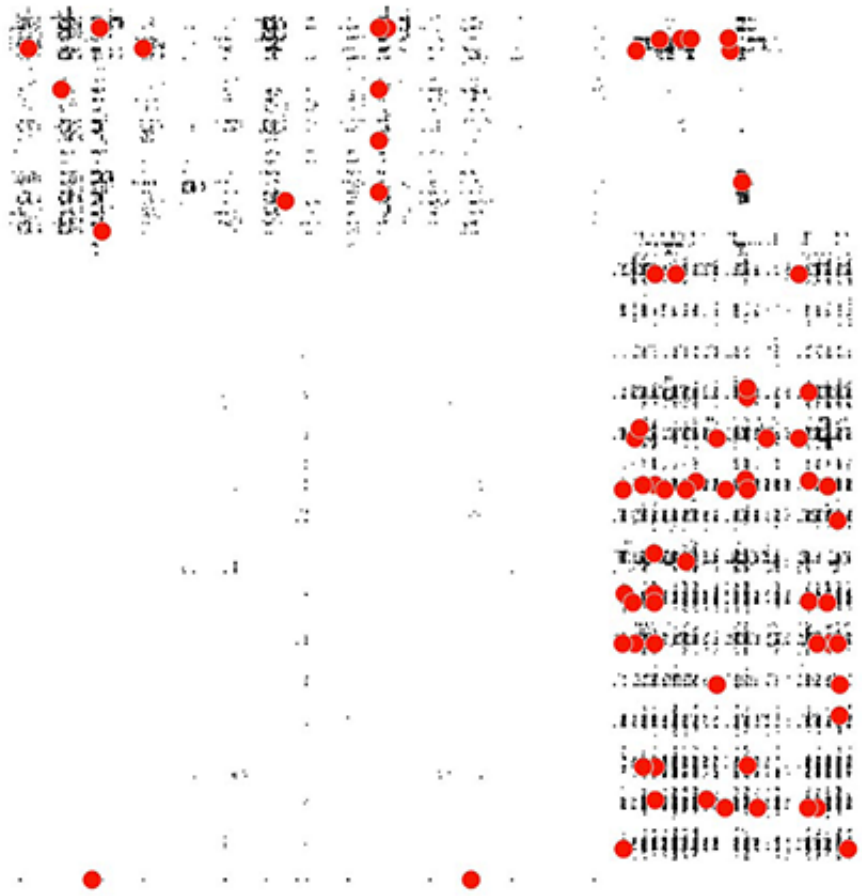

Figure 11. The text of John Donne's "O my black soul" sonnet visualized with English Phonemes.

6.0 QUALITATIVE STUDY

We wanted to discuss this project with a cross-section of scholars to develop a better understanding of how people understood L-DNA and whether or not they saw potential uses for their research. Our goal was to gain insight into whether this technique could inspire reactions and possibly spark interest in the approach. 58

6.1.1 Participants

We intentionally sought participants from a variety of disciplines. We had 14 participants which included 1 visual artist, 3 literary critics, 1 rhetorician, 2 digital media critics, 1 database programmer, 1 business analyst, 1 linguist, 2 interaction designers, 1 graphic designer, and 1 marketing specialist. 59

6.1.2 Procedure

Each participant took part in a thirty-minute interview and was first shown L-DNA visualizations that we had intentionally left void of any legends or axis labels, so that we could ask questions about their initial interpretations. After showing the image in Fig. 4, the mapping of the OED in two dimensions, we asked what they thought the image might be. We then showed participants Fig. 7, four languages plotted in the same space, and the interviewer gave a thorough explanation of the how the algorithm works and what they were seeing. We took time to make sure they were comfortable with the explanation and asked about their understanding. We then showed them Fig. 8 to be able to further explain L-DNA and asked questions based on the participants' understanding. We also asked participants to complete a post-interview questionnaire. Five questions were asked on a 5-point scale, with an opportunity to provide free form answers: 60

Q1. Once explained to what extent is the visualization readable? 61

Q2. Do you think the white space has meaning? 62

Q3. Is the white space necessary? 63

Q4. Does representing languages by colour and words as points work well? 64

Q5. Does this spatial representation of language trigger new ideas? 65

The following three questions asked for free form answers only: 66

Q6. What are your initial interpretations of this visualization? 67

Q7. Can you imagine a more suitable or readable structure? 68

Q8. Please provide any criticism you have about this visualization. 69

7.0 RESULTS: SCALE-BASED QUESTIONS

The scale questions were answered as follows. For Q1, 6 out of 14 participants told us the visualization was clear after the explanation (5 out of 5 on the scale). FOR Q2, 9 out of 14 people said that the white space carried meaning to them (5 out of 5) and 6 participants thought that this white space was completely necessary. For Q4, 7 out of 14 people ranked a 5 out of 5 for the visualization approach. 10 out of 14 participants said that the visualization triggered new ideas for research (5 out of 5). 70

8.0 DISCUSSION: FREE-FORM QUESTIONS & INTERVIEWS

We have formulated our discussion around the free-form questions. Since our participants were experts from a variety of fields and domains, the similarities in answers in some cases are particularly interesting. In other places it is the difference in answers that encourages us as researchers in terms of the potential of L-DNA as a tool for approaching questions about language. In this discussion we include the questions and a discussion of the general themes that arose in the responses. 71

8.1 The Power of Representation

Interestingly, when shown the images without any labels, participants tended to engage in metaphoric comparisons of what they were seeing. The omission of a legend led each participant to find something that was cognitively analogous to what they were being shown. For example, some responses were: 72

“Is it zoomable? It has a DNA look or sort of ummm a matrix data flow and I feel the urge to zoom it. It looks like I am really far away from an unbound book, like how a book would be printed in sheets. It has that type of aesthetic.” 73

“Well, it reminds me of DNA, like a screening test, it also looks like a stamp, like someone has stamped something. It is a very tactile image, I want to touch it.” 74

“It’s kind of like DNA. Like, uh, people show these images that visualize DNA.” 75

In response to this process, five out of nineteen (26%) of our participants from varying backgrounds and fields, with no explanation of what they were looking at, described Fig. 4 as appearing like DNA — the inspiration for the name of our technique. This result also demonstrates the power of representation held within L-DNA. Some of the responses received from participants included references to stamped or fading paper, city grids, abstract art, and digital clock faces. Because we were trying to create a space that could handle an ontology of words, the DNA metaphor was highly applicable based on the implications of describing parts of a long chain of information. 76

8.2 White Space

In our interpretation of the study data, the questions about whitespace (Q2-Q3) produced perhaps the most interesting 77

results. It was during this question that most of the participants began to hypothesize about the “space” in the languages they use every day. Essentially the parts of the visualization with an absence of marks inspired thought, because they were in stark contrast to the actual dots drawn on the screen. This result strongly indicates the analogic or comparative possibilities of this technique — the literary critic can begin to understand what makes a word English by recognizing what is not a word, or by investigating what words poets or writers use that push the boundaries of language. This result is encouraging for the domain-specific problems of literary criticism. One of our requirements is a space for analogues and with the whitespace in this simple mapping there is significant affect. The response of our participants to the relationship between the whitespace and the space occupied by words creates a relationship that gives insight into what sets of symbols we use and which we do not use in our language. The sheer size of the whitespace in comparison forces an understanding of how few of the possible arrangements of letters we actually use. With further work we think it is possible to show that more complex mappings will produce more complex analogues.

It was generally agreed upon that it was the relation of the empty space to the marked space that created meaning and inspired insight into what the visualizations were showing and what they could show. 78

“the white space gives you a sort of ground against which it makes sense... without it it might be even less evident” 79

It was in the white, or the lack of spellings, that our participants saw the potential for growth in the language, or commented on the enormous range of letter combinations that we do not use in English. This is encouraging because as we build “meaning” into these visualizations we expect the response to be comparative, and we expect new interpretations will result from these comparisons. Initial reasons were as follows: 80

“Every letter can start a word, but does that mean that there’s combinations of letters that don’t turn into words... you don’t have words that start with trsz is that why there would be space... Maybe it means that language is primitive, not as evolved as it could be.” 81

“I’m almost more interested in the dots in the white space... if there's one I want to know what that is...the outliers are more interesting.” 82

“What it does is show what the common letters are and common overlaps.” 83

For the researchers this result was surprising, but was explainable. Without “meaning” being built into this version we were simply showing a part of orthography (spelling) and, in this stripped down version of the possibilities of the space, it was the comparison between what was empty and what was marked that sparked the interest of our participants. This is the exact response we were interested in and we anticipate with more complex representations we will be able to see more complex analogies. Some of the responses to this whitespace analogy are as follows: 84

“I think the uniformity of the gaps are startling. It seems oddly uniform and consistent. I think that potentially symbolically the gaps can sort of be a formative quality of language and words will start to fill in those gaps.” 85

“Well, I guess it kind of goes to show how constrained we are in language, which shows how some things just are not possible with spelling, which is kind of cool. There is so much blank space particularly along certain lines, you get some sense that our alphabet constrains us, which is why we have poets.” 86

These responses were typical of all our participants and are very promising for future work. 87

8.3 Reluctant Inspiration

When asked if the images they were seeing inspired any new thought processes, they proved to be exciting to our participants and the answers that follow show the breadth of their thinking: 88

“You could map to any narrative, I would like to see this map out, A Tale of Two Cities, it looks like a genetic footprint, like a genetic phonetics, you get to see this formal genetic blueprint, it is more like an autopsy of the text.” 89

“Yes, from a design and a fine art point of view. I think it is, again, if you look at it as a design issue does it solve any 90

problems, not yet, but then again you are deconstructing something that already has that problem solved so you are raising questions instead of solving problems and you are raising interesting and meaningful questions."

"Well, I mean my initial instinct is YES (emphasis given by the participant), but I am not sure what that would be yet, I think it could lead to a lot of productive conversations about how language operates or the way someone creatively uses language." 91

"These are important questions. Literature departments have always performed as if they were in the shadows of the sciences and…[i]t seems to me that this kind of work, although seemingly scientific, should be the domain of literature department. It would be very important work for us to do." 92

Interestingly the 1 person out of 14 who said no to being inspired (included below) touched directly on the fact that the simplicity of our representation of orthography was limiting but suggested in his negative response that it could be interesting for literary criticism if we could find a way to include meaning, which is fruitful ground for future work and possible with the three criteria we laid out above for the development of the technique: 93

"I'm not sure it's useful for literary theory or criticism because it seems to explicitly set aside the question of meaning in favor of orthography" 94

"Oh yea, um this is only for spellings well things get really interesting when you get into phrases, rhetorical aspects of language organization, the whole question of nuance which we like to think we are studying as literary scholars. Once you get into these things and away from mere spelling and into points that have meanings you can start to clump together interpretation to different meanings." 95

8.4 Criticisms from Participants

Our participants were in some cases critical of the design of L-DNA. In particular, the most common criticism centered on making the technique interactive, which is a clear (and previously planned) next step in our iterative design process. Another criticism was that the encoding we used was arbitrary, and its meaning was not immediately clear: 96

"The mechanism by which you map words to spatial locations, it's kind of arbitrary (maybe that's not the right word). Mostly you see the first two letters, and when you zoom in you're seeing the space of subsequent letters. I guess what this mapping is lacking is the Meaning of words, or the semantics. They end up being distances from one another but … the distance between related concepts, synonyms. I'm not necessarily saying this is a bad visualization because of that. It just doesn't encode the meaning of words. Which is much harder. But, this is a legitimate way of putting words in a consistent space." 97

"The way that it is right now, that it's static, it's getting in the way of itself. It demands explanation. It needs to be paired up with something very practical, like reading a word or reading a sentence and how that gets paired up in the system." 98

We see this interactivity as the next steps in developing an application for these types of interactions and it is in that interactivity that the objections to the arbitrariness of the design will be addressed. We hypothesize that being able to investigate the space dynamically, and by defining other symbol systems to encode, the literary critic will be able to explore the types of meaning and associations being looked for by our participants. 99

9 STUDY RESPONSE

From our qualitative study, the overwhelming result was the importance of the white space in our visualizations to the entire group that was interviewed. This reaction has influenced the direction of our future work and demonstrates that this space can be used for the types of comparisons that we are interested in, namely those that lead to interpretive possibilities. We recognize that we have presented a simplified form, but the technique itself allows for infinite complexity. Some of our participants talked about the need to include information that gives meaning to relationships and that is the next step in developing mappings that can solve the original problem of creating a space that can be experimented in with language and literary theory. Our study confirmed the idea that this technique has the potential to 100

answer these much more complex questions as they relate to the domain in question. The sheer breadth of answers to our question in relation to inspiring new ideas is extremely promising and we take it as a success in developing the type of space that can inspire investigation.

9.1 Prioritizing Whitespace

In response to the discussion with participants about the interest in whitespace and the lack of density in certain regions, we produced a density and inverted density map to highlight the white spaces, shown in Fig. 12 and Fig. 13.

Figure 12. Density map of the English language in L-DNA.

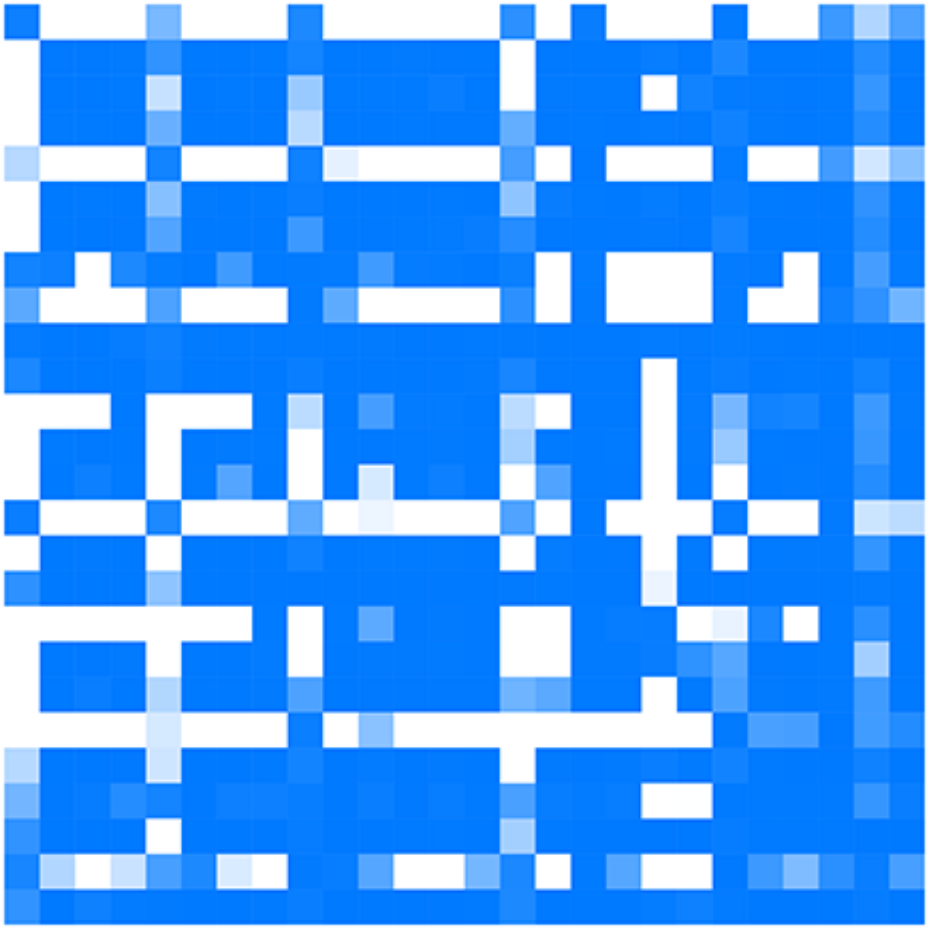

Figure 13. Inverted density map of the English language in L-DNA.

This was in direct response to comments from our participants such as: 102

“Yeah. But, one thing I was going to say… the UNIFORM inclusion of the whitespace is interesting. But I wonder if there’s a better representation of density and overlap.” 103

In this iteration of our design, instead of rendering each word in the language we instead cluster groups of words into the boxes representing pairs of words (“AA”, “AB”, etc.), and render the box using transparency that corresponds to the density of words therein. The inverted version of this mapping highlights all of the non-English words, which were clearly of interest to participants. 104

9.2 ygUDuh

We have discussed how the blank spaces are compelling and how the absence of words in these spaces generally seems to intrigue people. 105

It is interesting to note that this space is and has been filled in many interesting ways. For example, E. E. Cummings poem ygUDuh does not contain a single English word, yet it can be read as English, where the “words” take on the sonic characteristics of English when read aloud. For example, the first few lines of the poem are: 106

```
ygUDuh

      ydoan
      yunnuhstan
```

That when read allowed becomes a phonetic map for a type of early 20th century urban slang exemplified by the poem: 107

```
you gotta

      you don't
      you understand
```

In Fig. 14 we have plotted ygUDuh overlaid on the OED grid. Note how these "non words" exist largely on the edge of the word spaces and the white spaces. This may be because, while they are not English words — hence the white space proximity — they have similar vowel and consonant structures to English words. By seeing these words overlaid on the whole language, we can see a visual representation of Cummins' craft, of the attempt to make non English words that sound like English when read aloud. The fact that all of the non-words are situated on the edges of heavily populated space tells us that these arrangements of letters that we try to make into words when reading the poem are "closer" to English words than we think, at least in terms of spelling. 108

Figure 14. ygUDuh overlaid on the OED grid.

9.3 Instant Messaging

Another example where new types of "words" or at least English communications are evolving is in instant messaging, text messaging, and social media. It seems that for ease and speed, we can give up many letters — chiefly the vowels — in words and still retain meaning. Fig. 15 shows MSN "words" overlaid on the OED visualization. It is interesting to note that many of the new "words" (marked with red dots), fall in the spaces where very little or no words exist. This demonstrates that even in a type of shorthand, like the one used in instant messaging (e.g., "btw", "lol", "ttyl", etc.) that many of the newly created words are spelled with letter combinations that simply don't exist in the language. This is partially a result of the volume of acronyms used in instant messaging but it becomes obvious by "reading" the image that many of the words used fall on the top line of each row suggesting (such as with row A, and row I) that many of these "words" and acronyms begin with those letters. In this way our technique produces visuals that allow us to ask further questions about the organization of our data. 109

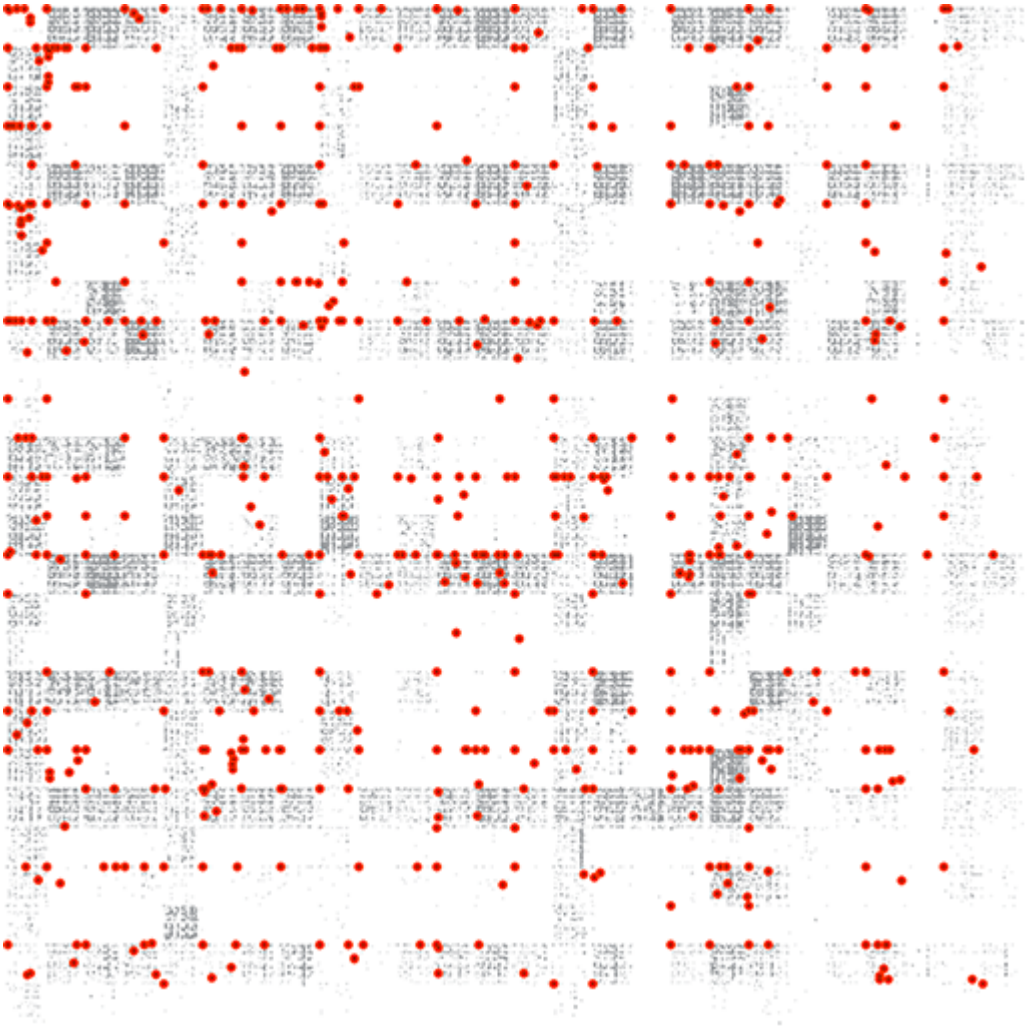

Figure 15. MSN ‘words’ overlaid on the OED visualization

9.4 Interaction

We have also begun to integrate interactive elements into our visualization, some of which were planned prior to our qualitative study, and some of which were inspired by our results. In particular, we have already created a version of L-DNA which incorporates a brushing technique that presents the “words beneath the cursor”, both when dragging across words and when dragging across the whitespace. We have also created a version of the density map (Section 9.1) that allows zooming into the recursive letter pairs. For example, it is possible to click or tap on the “BA” square, then the “NA” square, then another “NA” square to then see the word “BANANA” as shown in the static image of figure 3. 110

10 CONCLUSION

In this paper we have introduced L-DNA and presented the findings of a qualitative study of its design. In L-DNA, we have developed a mapping of symbol systems to visual space, which we have demonstrated using language. Our formulation has several properties that are valuable for the analysis of language and are not available in some other common visualizations of language. L-DNA has the following important features: 111

1. The L-DNA space is capable of handling any symbol string from a null string to a string of infinite length. L-DNA can be used in 1, 2, or n dimensions.
2. L-DNA space is infinite in that between any two points (words) in the space there exists another word — though it may not have semantic meaning.
3. The L-DNA space is one-to-one and onto (bijective). Every unique coding maps to a unique position and, in

reverse, words (or any original information) can be recovered uniquely from the visual space.

We have mathematized language to make exploring and experimenting with language easier, but the results of said experiments need to have the possibility of being reversed out of the space to be able to assign meaning once again to the language. We have also presented a qualitative study, which provided encouraging results that indicate the power of the type of representation provided by L-DNA, the benefit of the whitespace that it generates, and its possibilities to provide inspiration (even if reluctantly), as well as some useful criticisms that led to iterations in our design.

Notes

[1] https://books.google.com/ngrams

[2] We present only three simple possibilities here. It must be stressed that the algorithm works for any symbol system or combination of symbols systems.

[3] We have removed words with diacritics and punctuation for simplicity of demonstration. The algorithm is fully capable of handling these symbols with slight changes to the function.

Works Cited

Alexander 2012 Alexander M., “Patchworks and field-boundaries: visualizing the history of English”, Digital Humanities (2012), Hamburg Germany.

Bingenheimer 2006 Bingerheimer M., Xishihu, J. “Social network visualization from tei data”, Lit Linguist Computing 26 (3): 271-278 (2006).

Bingenheimer 2009 Bingenheimer, M., Hung, J., and Wiles, S. “Markup meets GIS – Visualizing the ‘Biographies of Eminent Buddhist Monks’”, In Banissi, E. et al. (eds), Proceedings of Information Visualization IV Piscataway/NJ: IEEE Computer Society, pp. 550–4 (2009).

Chen 2006 Chen, C. “CiteSpace II: Detecting and visualizing emerging trends and transient patterns in scientific literature”, JIS&T 57, 359–377 (2006).

Chuang 2012 Chuang J., Ramage D., Manning C., Heer, J. “Interpretation and Trust: Designing Model-Driven Visualizations for Text Analysis” ACM Human Factors in Computing Systems (CHI), 2012.

Clement 2008 Clement, T. E. “A thing not beginning and not ending: using digital tools to distant-read Gertrude Stein’s The Making of Americans.”, Literary and Linguistic Computing, 23(3): 361–81 (2008).

Collins 2009a Collins, C., Carpendale, S. and Penn, G. "Docuburst: Visualizing document content using language structure." Computer Graphics Forum. Vol. 28. No. 3. Blackwell Publishing Ltd (2009).

Collins 2009b Collins, C. Viegas, F. Wattenberg, M.. "Parallel tag clouds to explore and analyze faceted text corpora." Visual Analytics Science and Technology, 2009. VAST 2009. IEEE Symposium on. IEEE (2009).

DeCamp 2005 DeCamp, P., Frid-Jimenez, A., Guiness, J., & Roy, D. “Gist icons: Seeing meaning in large bodies of literature”, In *Proc. of IEEE Symp. on Information Visualization* (InfoVis 2005), Poster Session, Minneapolis, USA, October, (2005).

Dict 2015 http://www.winedt.org/Dict/ (Accessed April 19 2015)

Frye 2000 Frye, N., The Anatomy of Criticism, Princeton University Press (2000): 7.

Gardiner 2010 Gardner, M. J., Lutes, J., Lund, J., Hansen, J., Walker, D., Ringger, E., and Seppi, K. “The Topic Browser: An interactive tool for browsing topic models”. In NIPS (Workshop on Challenges of Data Vis), (2010).

Gould 2003 Gould, R. V. “Uses of network tools in comparative historical research”, in Mahoney, J. and Rueschemeyer, D. (eds), Comparative Historical Analysis in the Social Sciences. Cambridge: Cambridge University Press, pp. 241–69 (2003).

Gretarsson 2011 Gretarsson, B., O’Donovan, J., Bostandjiev, S., Llerer, T. H., Asuncion, A., Newman, D., and Smyth, P. “TopicNets: Visual analysis of large text corpora with topic modeling”, Trans on Intelligent System and Technology 3, 2 (2011).

Hall 2008 Hall, D., Jurafsky, D., and Manning, C. D. “Studying the history of ideas using topic models”. In EMNLP, 363–371 (2008).

Hearst 1995 Hearst, Marti A. “TileBars: visualization of term distribution information in full text information access.” *Proceedings of the SIGCHI conference on Human factors in computing systems*. ACM Press/Addison-Wesley Publishing Co., (1995).

Hetzler 1998 Hetzler, Beth, et al. “Multi-faceted insight through interoperable visual information analysis paradigms.” *Information Visualization*, 1998. Proceedings. IEEE Symposium on. IEEE, (1998).

Jessop 2006 Jessop, M. “Dynamic Maps in Humanities Computing”. Human IT, 8.3, 68–82 (2006).

Jessop 2008 Jessop, M. “The Inhibition of Geographical Information in Digital Humanities Scholarship”, Literary and Linguistic Computing, 23(1): 39– 50, (2008).

Lee 2010 Lee, Bongshin, et al. “Sparkclouds: Visualizing trends in tag clouds”. *Visualization and Computer Graphics, IEEE Transactions on 16.6*: 1182-1189, (2010).

Lin 1992 Lin, X. “Visualization for the document space”. In Vis, 274–281 (1992).

Moretti 2005 Moretti, F, Graphs, Maps, Trees: Abstract Models for a Literary Theory. London: Verso, 2005.

Paley 2002 Paley, W. Bradford. "TextArc: Showing word frequency and distribution in text." Poster presented at IEEE Symposium on Information Visualization. (2002).

Posavec 2015 http://www.stefanieposavec.co.uk/-everything-in-between/#/writing-without-words/ (Accessed April 19 2015)

Rohrdantz 2012 Rohrdantz, C., Niekler, A., Hautil, A., Butt, M., Keim, D. “Lexical semantics and distribution of suffixes: a visual analysis”. *EACL 2012 Proceedings of the EACL 2012 Joint Workshop of LINGVIS & UNCLH*, pages 7-15. (2012).

Simon 1988 Simon, Herbert A. “The science of design: creating the artificial.” *Design Issues*: 67-82,(1988).

Smalheiser 2009 Smalheiser, N. R., Torvik, V. I., and Zhou, W. “Arrowsmith two-node search interface: a tutorial on finding meaningful links between two disparate sets of articles in MEDLINE” Computer M&P in Biomedicine 94, 2, 190–197 (2009).

Tapor 2015 portal.tapor.ca (Accessed April 19 2015)

TextRain 2015 http://camilleutterback.com/projects/text-rain/(Accessed April 19 2015)

Valley 2015 Valley of the Shadow http://valley.vcdh.virginia.edu/ (Accessed April 19 2015)

Viegas 2008 Viegas, F. B., and Wattenberg, M. “TIMELINES: Tag clouds and the case for vernacular visualization. Interactions” 15, 49–52 (2008).

Viegas 2009 Viegas, Fernanda B., Martin Wattenberg, and Jonathan Feinberg. “Participatory visualization with wordle”. *Visualization and Computer Graphics, IEEE Transactions on 15.6*: 1137-1144, (2009).

Voyant 2015 http://voyant-tools.org/ (Accessed April 19 2015)

Wattenberg 2008 Wattenberg, Martin, and Fernanda B. Viégas. "The word tree, an interactive visual concordance." Visualization and Computer Graphics, IEEE Transactions on 14.6: 1221-1228 (2008).

Wise 1995 Wise, J., Thomas, J., Pennock, K., Lantrip, D., Pottier, M., Schur, A., and Crow, V. “Visualizing the non visual: spatial analysis and interaction with information from text documents”, in InfoVis, 51–58,(1995).

Wood 1992 Wood, D., The Power of Maps. New York: Guilford Press (1992).

WordCollider 2015 http://www.itsokaytobesmart.com/post/26993944713/word-colliderparticiple- physics (Accessed April 19 2015)

Yatani 2011 Yatani, K., Novati, M., Trusty, A., & Truong, K. N. “Review spotlight: a user interface for summarizing user-generated reviews using adjective- noun word pairs”. In Proceedings of the 2011 annual conference on Human factors in computing systems (pp. 1541-1550). ACM (2011).

van Ham 2009 van Ham, F., Wattenberg, M., and Viegas, F. B. “Mapping text with phrase nets”, in InfoVis, 1169–1176 (2009).