CARESSING THE NOUN: USING NATURAL LANGUAGE PROCESSING TO READ GERTRUDE STEIN’S TENDER BUTTONS

A Thesis
submitted to the Faculty of the
Graduate School of Arts and Sciences
of Georgetown University
in partial fulfillment of the requirements for the
degree of
Master of Arts
in English

By

Emperatriz Kathleen Ung, B.A.

Washington, DC
April 3, 2017
ABSTRACT

This thesis uses Natural Language Processing to study Gertrude Stein’s *Tender Buttons*. Stein’s poetics collide with alternative reading practices. Her writing challenges both conventional and nonconventional approaches, and disrupts the “seeing” of machine reading that utilizes Natural Language Processing. Her concern with grammar and vocabulary are rendered visible with computer-aided reading, revealing patterns that reinforce her commentary in “Poetry and Grammar.” Stein’s writing demands special attention in traversing the balance between traditional, close reading and computational approaches.
The research and writing of this thesis is dedicated to Darian Li and everyone who supported me on this journey.

Many thanks,
Emperatriz
# TABLE OF CONTENTS

INTRODUCTION .......................................................................................................................1

PART I. HOW DIGITAL TOOLS HAVE BEEN USED TO STUDY LITERATURE .......... 3
  How Scholars Use Close Reading .........................................................................................3
  Emergence of Machine Reading ...........................................................................................4
  Examples of Machine Reading Twentieth-Century Texts ..................................................10
  Machine Reading in Part II—Natural Language Processing ..............................................18

PART II. MACHINE READING GERTRUDE STEIN .........................................................23
  Previous Digital Approaches to Stein ..................................................................................23
  The Difficulty with Reading Stein .......................................................................................25
  Parts of Speech in Stein’s Work ..........................................................................................30
  Reading Stein with NLP Data .............................................................................................36

CONCLUSION .........................................................................................................................45

APPENDIX I .............................................................................................................................47

APPENDIX II ...........................................................................................................................50

APPENDIX III ..........................................................................................................................52

APPENDIX IV ..........................................................................................................................54

BIBLIOGRAPHY ......................................................................................................................56
Introduction

This thesis started as a creative experiment in response to Gertrude Stein’s essay “Poetry and Grammar.” As a poet and scholar, I was curious about Stein’s claim that poetry was about caressing and addressing the noun:

Poetry is concerned with using with abusing, with losing with wanting, with denying with avoiding with adoring with replacing the noun. … When I said. A rose is a rose is a rose is a rose. And then later made that into a ring I made poetry and what did I do I caressed and completely caressed and addressed a noun (LIA 231).

She writes with much urgency. Stein’s essay suggests nouns contain poetic power that changes across genres, and that poetry in particular was unique in its relationship to parts of speech. My creative response took the form of generative art as a Twitter bot. Questions that guided the original project included: Does Stein's writing still maintain a poetic “ring” if the nouns are in a constant shifting state? Does the regeneration of poetic lines cause instability in the reading of a work? How does this all translate when the pattern is placed in a programming environment? A collection of nouns was selected from within her original writing. Each word of Stein's “Sacred Emily” was searched in the Oxford English Dictionary, and words were identified as nouns based on the OED entries. Rare and obsolete entries for nouns were excluded. I grouped nouns by the number of syllables to try and maintain the rhythm of Stein’s original lines. Syllables acted as “beats” to adopt qualities of quantitative verse.

This project and its questions later transformed into a scholarly curiosity after conversations I had with linguists and data scientists. In what other ways could we study
Stein’s use of nouns? As the project became more scholastic I began searching for digital tools that would complement my research questions in order to engage with current digital approaches to the study of literature.

Scholars such as N. Katherine Hayles, Franco Moretti, and John Guillory have all examined critical reading practices in literary study. For computational literary and cultural studies in particular, Franco Moretti’s distant reading approach has emerged as a dominant methodology in favor of big data. This thesis builds on the work of these scholars to argue for the value of machine reading at a local level—not only for quantitative and statistical claims, but also to supplement close reading analyses. Difficult modernist texts, such as Gertrude Stein’s *Tender Buttons*, thwart the distant reading approach outlined by Franco Moretti. Not only does Stein’s formally innovative writing frustrate this computational method of literary study, but her writing, too, subverts Moretti’s claim of literature as world system that can be aggregated on a global level.

This thesis is split in two parts. The first part frames current research in digital humanities, digital approaches to twentieth-century Anglophone literature, and the computational tools used in this thesis. The second part analyzes Gertrude Stein’s *Tender Buttons* with Natural Language Processing. The appendices contain the Natural Language Processing data collected at earlier stages of this study.
Part I. How Digital Tools Have Been Used To Study Literature

How can different reading approaches offer fresh insight to literary texts? The relationship between literary studies and close reading demonstrates how the introduction of machine reading supplements analyses of literature. This section explores how scholars read, outlines the literary landscape of digital approaches to the study of literature, and introduces how scholars have used machine reading. This section will also look at particular studies that use digital tools to study twentieth-century literature in order to frame the following study in Part II on Gertrude Stein.

How Scholars Use Close Reading

Many scholars rely heavily on the methodology of close reading. Jane Gallop calls close reading “the habit of literary reading” and offers an overview of how the practice of close reading became valuable to the discipline:

According to the standard histories of English, when the New Critics introduced the methodology called close reading, what it replaced was literary history (the old historicism, we might call it). We became a discipline, so the story goes, when we stopped being amateur historians and became instead painstaking close readers. I would argue that the most valuable thing English ever had to offer was the very thing that made us a discipline, that transformed us from cultured gentlemen into a profession. Not because close reading is necessarily the best way to read literature but because close reading--learned through practice with literary
texts, learned in literature classes— is a widely applicable skill, of real value to students as well as to scholars in other disciplines (Gallop 15).

Close reading today is a skill applied in other fields, including cultural studies. The act of close reading is taught in English departments to push writing beyond the “old historicism.” Gallop doesn’t offer alternatives to close reading, but is mostly concerned with how close reading is currently being used in the discipline of English literary studies. Over time the “painstaking” process of close reading has become synonymous with critical reading. Gallop provides theses on critical thinking and close reading in order to establish how close reading is specifically a disciplinary contribution to the study of literature, language, and rhetoric rather than a disciplinary contribution to the study of philosophy and ideas (18). For the purposes of this project, close reading is a type of critical reading.

Emergence of Machine Reading

As the field of digital humanities emerges, scholars work to theorize how technological developments might influence research practices in literary studies. The utility of digital tools can expedite analysis of large volumes of texts and otherwise support or debunk readings of a text. Ted Underwood discusses this utility as an asset in “revealing patterns of association we might otherwise overlook”: “But in fact I have used algorithms to explore a big dataset, and the search process may well have shaped my way of framing the subject, or my intuitions of the representativeness of sources” (Underwood 65). It shapes how we frame a thesis and gets us to reevaluate existing research practices.
A digital research method does not have to become “invisibly naturalized” to be useful in literary reading. Matthew Kirchenbaum and Sarah Werner discuss the importance of understanding the digital tools we use as literary scholars:

Not only do we need to learn what tools to take advantage of, the rest of the scholarly and public world needs our insights as part of the conversation, especially as the means by which information circulates today continues to shift in response to the technological and societal shifts around us (Kirschenbaum and Werner 410).

Scholars need to continue working toward understanding digital tools “invisibly naturalized,” widely used, or not frequently used.

Computer-aided reading, also known as machine reading, has become a computational method of literary and cultural studies. In N. Katherine Hayles’s article “How We Read: Close, Hyper, Machine,” Hayles fleshes out the scope of machine reading as practiced by scholars: “Machine reading ranges from algorithms for word-frequency counts to more sophisticated programs that find and compare phrases, identify topic clusters, and are capable of learning” (72). Human interpretation plays a critical role in machine reading. With the wide range of computer software and applications available to scholars, human interpretation shapes the stakes and claims formulated by data alternative reading approaches provide.

The role of close reading, according to Hayles, stems from an anxiety in the field when scholars turned to reading different media as texts in the 1970s and 1980s, such as film and architecture: “Faced with the loss of this traditional center, literary scholars found a replacement in close reading, the one thing virtually all literary scholars know
how to do well and agree is important. Close reading then assumed a preeminent role as the essence of the disciplinary identity” (63). To part with close reading jeopardizes this “essence of the disciplinary identity” in literary studies.

Looking to other fields, such as human-computer interaction and social sciences, the humanities can use emerging technology to engage in a dialogue beyond data mining of large collections. Pairing literary texts with a specific digital tool places them in conversation with one another rather than limiting the dialogue. Literary texts can inform the work and limitations of machine reading. Equal attention to the analysis and software leads to more comprehensive scholarship. Digital humanities and machine reading do not overshadow traditional modes of analysis, but instead become tools that extend the work of current humanistic research and “do not remove the need for human reading of the material” (Drouin 130). John Guillery writes:

Our understanding of how scholars read, then, depends on acknowledging the divergence between scholarly and lay modes of reading. This divergence was correlated with a diversification in genres of writing, but it would be a mistake to identify scholarly modes of reading as simply the effect of new genres of scholarly writing (9).

Guillery is quick to limit the value of machine reading, reducing it to the less than labor-intensive work on par with the “distortions produced by scanning” a text. He presumes that alternative reading approaches are lay modes of reading. The labor a scholar dedicates in designing a computational approach still requires interpretation of the findings. The “when and how” of deceleration continues to be critical to generating an analytical reading. The “distortions” and errors that arise from computer-aided reading
should not be ignored: “Yet, the main reason to report failures has nothing to do with
capturing the benevolence of readers; it’s that failures throw a unique light on the whole
research process” (“LM,” Moretti 4).

One trending form of machine reading is distant reading. Distant reading,
introduced to literary scholars by Franco Moretti, was formulated as a non-computational
approach to make larger cultural claims on literary trends. Born from structural and
quantitative analysis, Moretti proposes analyzing texts in opposition to close reading,
where close reading practices focus on traditional modes of analysis and are limited from
the outset. He contends, “we know how to read texts, now let’s learn how not to read
them” (DR, Moretti 48). His distant reading project responds to the problem of “the Great
Unread” in Weltliteratur, and is concerned with the volume of work that goes unread and
missing from scholarship (DR, Moretti 180).

For Moretti, dependency on a small literary canon hinders the way scholars read
and limits the scholarship produced. In “Conjectures on World Literature,” Moretti
developed distant reading as a means to aggregate information and synthesize the work of
other scholars without the aid of a computer. He expands this project to use
computational approaches in “Style, Inc.: A Reflection on 7,000 Titles,” where he uses
quantitative stylistics to investigate seven thousand British novel titles over a hundred ten
year span from 1740 to 1850. Rather than analyzing these novels in their entirety, the
titles of the novels become the element evaluated by machine reading. Moretti zooms out
to distance himself from the content of the texts. The unread novels remain unread in an
attempt to move toward scholarship that considers literature as a whole, not just canon.
This exposes the parameters in Moretti’s methodology—distant reading and machine
reading rework the unit, or object, being read. The work of reading a primary text no longer maintains the same implications:

Distant reading: where distance, let me repeat it, *is a condition of knowledge*: it allows you to focus on units that are much smaller or much larger than the text: devices, themes, tropes—or genres and systems. And if, between the very small and the very large, the text itself disappears, well, it is one of those cases when one can justifiably say, *Less is more.* If we want to understand the system in its entirety, we must accept losing something (*DR*, Moretti 48-9).

In altering the units one focuses on, the scholar loses what is commonly perceived to be the primary facet of the text. This is where machine reading and distant reading intervene with the practice of close reading. Moretti’s call for distant reading is at odds with the tradition of close reading, where close reading holds a highly valued role in literary scholarship. Both machine reading and distant reading still require close reading, but the focus of close reading shifts from the primary text to the designated unit of study. This distance from the text in “distant reading” is misconstrued as a replacement or substitute for close reading. Instead, distant reading is an example of the utility of machine reading. Moretti’s framing of distant reading and other computer-aided approaches have contributed to the growing volume of scholarship in computational literary and cultural studies.

The automatic reactions to these claims tend to ask for the benefits. Each text demands a different approach. There are complications with all methods. But close reading offers too narrow of a perspective, and doesn’t scan a text or body of work uniformly. Scholars such as Julie Orlemanski have discussed the relationship between
close reading and distant reading. She primarily engages with Moretti’s devaluation of close reading and specifies how scale factors into the distant reading approach.

Orlemanski writes:

Moretti assembles texts without recourse to close reading’s interface, without recursive oscillation between text and context, form and history. In other words, he steps out of the hermeneutic circle. His peeling apart of the categories of the “text” and of the “proper object of knowledge” is the crux of the scalar challenge to literary study (223).

What Orlemanski observes here is an underlying current that drives Moretti’s projects—a resistance to traditional approaches. Orlemanski, too, highlights how this is not restricted to one particular literature and how units that become the objects of study are not static.

Working with the misconception that machine reading works against close reading, this thesis aims to reconcile how reading practices can be combined and why this combination is useful to literary scholars. Moretti develops the distant reading approach to aggregate big data on large volumes of literary texts. Distant reading approaches, as framed by Moretti, primarily function to create distance from a text by “providing an essential—still missing—piece to the computational analysis of literature” for different units, including plot, titles, and genres (DR Moretti 211). Moretti is right to tailor his distant reading approach to these different literary units, but his approach is misleading in suggesting that an “essential” framework can be achieved with machine reading. Through a variety of literary questions he looks to address, Moretti embraces “(mis-)using network theory to bring some order into literary evidence, but leaving my analysis free to follow
any course that happened to suggest itself” (212). His experiment with machine reading opens opportunity for other scholars to engage with these alternative practices.

**Examples of Machine Reading Twentieth-Century Texts**

Machine reading has been used to study twentieth-century Anglophone literature. Some examples of machine reading modernist works involve variations of distant reading. Hoyt Long and Richard Jean So shift between close reading a poem’s form, analyzing a poem as a sociohistorical event, and machine learning to ultimately contribute to the history of modernist Orientalism in the United States. Long and So suggest this combination to synthesize different approaches leads to new contributions: “Here we move beyond this impasse by modeling a form of literary analysis that, rather than leveraging one mode of reading against another, synthesizes humanistic and computational approaches into what we call literary pattern recognition” (Long and So 235-6). Rather than favor one approach, Long and So recognize that a balance between approaches is necessary to advance study.

Long and So use machine learning as a method of machine reading. For Long and So, machine learning “refers to a whole suite of statistical algorithms that treat every text as an amalgam of certain quantifiable features” (Long and So 250). They trace the use of machine learning in literary studies and attribute it to earlier manual efforts of information filtering. Much like Franco Moretti’s earlier work in “Conjectures on World Literature,” machine learning stems from earlier techniques that did not depend on computers to lay the groundwork for pattern searching.
They first look at their corpora as objects of close reading in order to establish parameters for their algorithm (Long and So 242). Close reading is involved in both the early stages and later stages of the study. This requires labor beforehand in order to initialize the study. While manual efforts are key to successful machine reading studies, the role of machines offers an angle not readily available with just close reading:

From the perspective of close reading alone, the poem does not rigorously fulfill certain criteria as laid out by scholars…. Machine learning, on the other hand, suggests that there is some relation to haiku at the level of statistical pattern—a subtle yet consistently present pattern of words and collocations of words. This influence as a kind of statistical likelihood, where words and other stylistic features are seen to be uniquely distributed across different types of texts. These latent, nonexplicit traces of influence are precisely what the machine is good at detecting and are impossible for the individual reader to identify on a large scale (Long and So 266).

Their example highlights the advantage of pairing machine reading with close reading. In this literary pattern study, Long and So have determined that close reading misses the “nonexplicit” aspects of the poems that machine learning detected. Close reading and machine reading are dependent on each other in order to produce a thoughtful study. The nature of designing a project with machine reading influences how the machine will respond to the text: “And yet it is important to remember that even in these cases where the machine’s determination is not aligned with that close reading or cultural history might tell us, these latter methods have informed the machine’s decision from the start”
(Long and So 266). The design of a study is crucial to how the machine will process the text.

Long and So also elaborate on their approach to combine close reading and machine learning. They reach the conclusion that this synthesis is critical to future work in the discipline in order to continue the work they have begun with their study: “…what we want to make clear is that it will require a method of reading that oscillates or pivots between human and machine interpretation, each proving feedback to the other in the critic’s effort to extract meaning from texts” (Long and So 267). This oscillation with human interpretation is important to the process of machine reading. It is this oscillation that underscores the imperative to better understand current reading practices as scholars. This project on Gertrude Stein looks to engage with Long and So’s approach to work between close reading and machine reading.

Like Orlemanski’s “Scales of Reading,” Jeffrey Drouin looks at different measures of reading in his article “Close- and Distant-Reading Modernism: Network Analysis, Text Mining, and Teaching The Little Review.” Drouin seeks the “middle ground” between human and automated techniques, to develop methods that “combine the macro and micro” (111). Quantitative analyses alone lack interpretation: “The main weakness of big data methodologies is their inability to read the works. The microscopic approach of text mining presents similar benefits and drawbacks” (Drouin 110). Big data methodologies overlook microscopic details in order to reveal other trends and patterns in a text. The process can lead to new directions in studying a writer’s oeuvre or larger collections. Drouin’s study looks at periodicals with different authors in order to analyze a particular historical moment. The combination of microscopic and macroscopic
approaches enabled him to situate his analysis in a more specific context. Drouin attempts to address the weakness of microscopic and macroscopic analyses with using both techniques in his study. His study depends on the human markup primarily provided by students over the course of four years. His approach to find the middle ground has taken years to get the corpus marked up: “More importantly, students write a brief description and assign topic tags to each item—such as World War I, Gender, or Death—in order to provide a sense of its meaning” (Drouin 119). The project depends on the labor of students and the data took years before becoming readily available for analysis. This approach required the effort of many:

…I hope to show the potential for a hybrid of human markup and automated analysis that would enhance our ability to understand the archive of modernist print culture. It turns on pedagogical activities in bibliographic description, semantic markup, and informatics that not only allow students to do real and valuable groundwork in laying bare the networks of modernism, but also serve as a model for more efficient use of stages in the research process (Drouin 116).

Here human interpretation is used at multiple stages of the process. Drouin turns to the data after students have curated content. Attention is given to the periodicals at a microscopic level. Students provide the bibliographic description that is used. This outsourced data becomes available for other scholars to use and perform further studies. This demonstrates the benefits of larger collaborative works with machine reading. The process leading up to machine reading can serve as a pedagogical tool. Machine reading requires thorough groundwork before initiating the process. Drouin’s model necessitates hours of groundwork beyond the work of a single scholar, but machine reading does not
have to only work on such a large, collaborative effort. Drouin’s process and method is not viable for a scholar to work alone. Many digital humanities projects require teams with members from different disciplines. This can be a challenge for a scholar working with limited resources.

Literary scholars using machine reading should be able to operate close reading along with interpreting quantitative results: “Effective digital literary study requires the ability to process data, read well, and interpret both the numbers and the texts in light of each other” (Drouin 111). Quantitative results still require human interpretation. This human interpretation becomes a close, or critical, reading of the data. Close reading and machine reading are not interchangeable. It happens one after another, as well as concurrently. Moretti, Long, So, and Drouin demonstrate how close reading occurs at earlier stages of a computational study, as well as how human interpretation contributes to existing scholarship. Drouin asserts, “Machines are not replacing the human factor in meaning making, but they can help us to look wider and delve deeper” (Drouin 130).

Machine reading can direct scholars to revisit past avenues of study with fresh attention. The labor intense work of close reading exists on both microscopic and macroscopic levels of machine reading practices. Human interpretation is important in order to not lose meaning in the research. This happens in other disciplines as well, where quantitative information alone does not provide meaning: “We can count, but we are rapidly forgetting how to say what is worth counting and why” (Weizenbaum 16).

Articulating the value of the numbers is an important part of a study.

Digital tools allow us to uncover information that human interpretation alone can sometimes overlook. Long and So emphasize this in their study: “Machines help us find
the patterns of relation that we have always known to operate in the creation and
diffusion of literary styles, but which until now we have been limited in our capacity to
recognize” (Long and So 267). Moretti also emphasizes this when he uses distant reading
to analyze the titles of novels in “Style, Inc.: A Reflection on 7,000 Titles.” The machine
reading reveals a trend that couldn’t be seen at the microscopic level.

Scholars can use computers as instruments in their studies: “The stories of man
and of his machine are inseparably woven together. Machines have enabled man to
transform his physical environment. … His tools, whatever their primary practical
function, are necessarily also pedagogical instruments” (Weizenbaum 17-18). Drouin
practices this with his students who markup the periodicals. Digital tools work with the
scholar’s critical reading, and machine reading does not force the scholar into lay modes
of reading. Machines and digital tools have been present in the academic work of
scholars for decades, as Underwood discusses in “Theorizing Research Practices We
Forgot to Theorize Twenty Years Ago” with the “invisibly naturalized” practices of data
mining (Underwood 64). Understanding the use of the machine, or digital tool, is
important to understanding how a particular tool can be used: “To understand the
interpretive limitations of an algorithm, one needs to understand its mathematical basis”
(Underwood 69). Underwood’s example of the algorithm highlights the strengths and
weaknesses of that approach. While the algorithm falls short of providing “authoritative
answers about the unity of discourse,” it succeeds at “revealing patterns of association we
might otherwise overlook” (Underwood 69). Algorithms, one type of digital tool,
advances studies like Drouin’s as well as Long and So’s.
Tanya Clement’s study on *The Making of Americans* is an example of a single work analysis. Rather than a corpus consisting many works, Clement focuses on a single book. The macro level of her study is *The Making of Americans*. She pairs close reading with machine reading in order to produce her study “The Story of one: Narrative and Composition in Stein’s *The Making of Americans*.” In her second study, she is able to synthesize her findings from distant reading the text:

In order to make a new reading of the text, I will use a critical reading with digital methodologies to show how Stein combines narrative (story and plot) from the first half of the text with non-narrative elements (such as accumulated, repetitive grammatical structures) in the second half of the text to create a mimetic experience of history in the making (“Story,” Clement 428).

Clement uses “critical reading with digital methodologies” in order to distant read Stein’s *The Making of Americans*. Labor-intensive critical reading skills are used to make sense of the information procured with digital tools. Machine reading comes after a close reading of the text. Clement’s approach to machine reading forces close, critical reading to be of value first. Her reading counters postmodernist readings:

Reading *The Making of Americans* as a mimetic, modernist text rather than a postmodernist text hinges on the discovery that Stein changes her patterns of repetition in the second half of the text. Though many have argued otherwise, there is evidence that compositional structure was very much part of Stein’s project for the meaning-making aspects of the text, and digital tools help us discover and read these elements in new ways (“Story,” Clement 436).
Clement’s project takes various steps to read and re-read *The Making of Americans*. She began with identifying misreadings in her earlier study, “‘A Thing Not Beginning and Not Ending’: Using Digital Tools to Distant-Read Gertrude Stein’s *The Making of Americans.*” This article that predates “The Story of one: Narrative and Composition in Stein’s *The Making of Americans*” details the process of distant reading the entirety of *The Making of Americans*. Clement claims that missed patterns become discoverable through the MONK (metadata offer new knowledge) project’s FeatureLens tool.

It is critical for scholars who practice machine reading to familiarize themselves with perspectives outside the discipline of literary studies. Long and So work to understand the network they have created by being aware of how their choices influence how the machine interprets the corpora. Lydia Liu attempts to bridge the work of the humanist and the work of the scientist:

I am not saying this to put the burden on the humanist and ask him or her to acquire a specialist’s knowledge and become a computer scientist. What I hope to emphasize is that there are good reasons for the humanist to get acquainted at least with what the scientist has done with our concepts of language, writing, and symbolic code at a level that goes deeper than the instrumentalization of reason (Liu 252).

Specialization is not necessary, but knowledge of the mechanics and basics of machine reading help the humanist better practice computational approaches to literary study. Close reading at multiple stages is part of the process.
How do scholars navigate the changing landscape of literary studies? With machine reading, the methodology must match the object of study. Machine reading is not one homogenous form or approach. There are a variety of digital tools that can be used for machine reading. The study that follows uses the Natural Language Toolkit with Python (NLTK). NLTK serves as a platform that uses the Python programming language to work with natural language data. Documentation of Natural Language Processing with Python defines natural language as: “language that is used for everyday communication by humans; languages like English, Hindi or Portuguese” (Bird et. al. n. pag.). Artificial languages are defined as “programming languages and mathematical notations” (Bird et. al. n. pag.). Natural Language Processing (NLP) is primarily used in computational linguistics, artificial intelligence, and human computer interaction. NLP functions for computers to understand natural language (Chopra n. pag.). NLP is part of machine translation, question answering, as well as information retrieval and extraction. Overall limitations to what NLP can help computers understand arise with real-world application. Programmers are confronted with these challenges when building natural language systems. These limitations are a problem primarily in the development of artificial intelligence:

Despite the research-led advances in tasks like RTE\(^1\), natural language systems that have been deployed for real-world applications still cannot perform common-sense reasoning or draw on world knowledge in a general and robust manner. We can wait for these difficult artificial intelligence problems to be solved, but in the

\(^1\) Recognizing Textual Entailment
meantime it is necessary to live with some severe limitations on the reasoning and knowledge capabilities of natural language systems. Accordingly, right from the beginning, an important goal of NLP research has been to make progress on the difficult task of building technologies that "understand language," using superficial yet powerful techniques instead of unrestricted knowledge and reasoning capabilities (Bird et. al. n. pag.).

The work of NLP is to foster communication between humans and computers. While the limitations on reasoning and knowledge capabilities prevent computers from seamlessly conversing with humans, NLP bridges the artificial languages to the natural languages.

Machine reading with NLTK allows the reader to control distance and scale. Tagged texts require both human interpretation and close reading in order to piece together quantitative results. This combination supplements how alternative reading practices factor into larger conversations concerning digital approaches to studying literature. Specifically, the part-of-speech tagging (POS-tagging) function of NLTK was used for the study that follows.

For NLTK to tag parts-of-speech, the corpus needs to be tokenized. Tokenization splits the text so that each word becomes a token. The tokens are then tagged for parts-of-speech. Tokenization is often the first step in NLP. In artificial intelligence projects, POS-tagging is merely the second step in bridging natural language with artificial language:

One of the earliest steps within this sequence is part-of-speech (POS) tagging. It is normally a sentence based approach and given a sentence formed of a sequence of words, POS tagging tries to label (tag) each word with its correct part of speech.
(also named word category, word class, or lexical category). This process can be labeled as a simplified form (or a subprocess) of morphological analysis. Whereas morphological analysis involves finding the internal structure of a word (root form, affixes, etc.), POS tagging only deals with assigning a POS tag to the given surface form word (Güngör 205).

Words become units within a sentence. Tagging the words into word categories assists the programmer to develop a set of rules for the computer to respond with in natural language. The difficulty with POS-tagging is that these categories shift between different linguistic theories and different natural languages. For example, Indo-European languages tag differently than East-Asian languages where certain semantic behavior is nonexistent. Distinguishing between different semantic behaviors is critical for researchers to later develop artificial intelligent programs that understand and respond in natural languages. Ultimately, the differing subcategories of linguistic theories are not crucial to the POS-tagging function of NLP: “Although the decision about the size and the contents of the tagset (the set of POS tags) is still linguistically oriented, the idea is providing distinct parts of speech for all classes of words having distinct grammatical behavior, rather than arriving at a classification that is in support of a particular linguistic theory” (Güngör 206).

Tunga Güngör identifies two main difficulties with POS-tagging systems—ambiguous words and unknown words. Computers have difficulty understanding how to process the information with ambiguous and unknown words. Researchers navigate these difficulties through different parts-of-speech tagging approaches. These different approaches consist of varying tagging algorithms and are based on different corpora or
linguistic theories. Some are rule-based approaches that require a linguistic background; other approaches include transformation-based learning (TBL), stochastic models, Markov (MM) models, and maximum entropy (ME) models. The default approach in NLTK is the ME model and used in this project. The ME model is more flexible than the MM and TBL model as it leverages “overlapping and interdependent features” to better use surrounding words as context in determining a word category (Güngör 219).

This project investigates how Gertrude Stein uses parts of speech in her poetic writing. *Tender Buttons* becomes the main corpora for the POS-tagging function. Understanding how the POS-tagging function works and where it comes from is necessary in understanding why it matters how literature conflicts with current NLP:

Ultimately software studies, critical code studies, and platform studies are each varyingly inflected methodologies for cultivating both the critical sensibility and the technical acumen necessary to swim deep into the cultural reservoirs of contemporary digital production, if not quite touch that final rocky bottom (Kirschenbaum and Werner 435).

Studying Stein not only shows how Stein’s writing resists being categorized in a traditional manner, but also how NLP’s strict algorithms have yet to allow room for intentional errors.

With Stein’s essays we can come to expect a certain style to emerge. On reading, Gallop notes that we must push past expectations to reach new conclusions:

When we read, we tend to see what we expect to have been written—what we expect *that* author to write, what we expect an author *like that* to write, what we expect from *that sort* of book. Reading what we expect to find means finding
what we already know; learning, on the other hand, means coming to know something we did not know before (Gallop 16).

The surprises come from close reading the output of NLTK. The search for the middle ground between human interpretation and machine reading is far from achieved. The efforts of this thesis works to use digital tools to provide quantitative information on Gertrude Stein’s oeuvre.
Part II. Machine Reading Gertrude Stein

In this study I combined close reading and machine reading to study Gertrude Stein’s *Tender Buttons*. Application of the Natural Language Toolkit with Python (NLTK) has been used as a form of machine reading to tag the primary text for parts of speech. With computer-aided reading, I isolated passages to highlight how experimental writing intersects and interferes with distant reading. Stein’s writing resists being tagged by the POS-tag NLTK function. Natural Language Processing (NLP) does not interpret some literature. An algorithm for literary studies and NLP does not exist. Such an algorithm would introduce literature to computers and in turn would allow for computers to author a unique genre of literature. This study on how Stein’s writing interacts with NLP informs both Stein studies and NLP studies. This section looks at previous studies of Stein’s work, the difficulty of reading Stein, Stein’s essays on writing, and analyzes *Tender Buttons*.

*Previous Digital Approaches to Stein*

Despite machine reading having some “distance” from the content of a text, this computer-aided reading delves into how Stein’s unconventional writing challenges both traditional and alternative reading practices. Tanya Clement utilizes Moretti’s approach to frame her study of Stein’s *The Making of Americans*. Clement suggests that the unreadability is due to “misreading” Stein’s difficult text (“Thing,” Clement 362). Clement writes:
The particular reading difficulties engendered by the complicated repetition in *The Making of Americans* by Gertrude Stein make it almost impossible to read this text in a traditional, linear manner. However, by visualizing certain patterns and looking at the text “from a distance” through textual analytics and visualizations, we are enabled to make readings that were formerly inhibited ("Thing," Clement 361).

Clement argues that “traditional, linear” readings prohibit scholars from accessing patterns, and that alternative reading practices give fresh attention to poetic elements of Stein’s writing. With the use of NLTK tagging Stein’s *Tender Buttons*, this project looks to further investigate how alternative reading practices relate to Stein’s own concerns with grammar and semiotics. Clement dedicates an entire article to her original findings in “A Thing Not Beginning and Not Ending: Using Digital Tools to Distant-Read Gertrude Stein’s *The Making of Americans.*” In “The Story of one: Narrative and Composition in Gertrude Stein’s *The Making of Americans,*” Clement synthesizes the work she did in her earlier study to do a critical reading of *The Making of Americans.* Clement’s study develops over the span of two articles. Understanding how previous digital approaches read Stein have functioned allows us to imagine future projects that intersect with the readability of Stein’s work.

Natalia Cecire focuses on the “unreadability” of Stein’s text, looking to previous scholarship that analyses Stein’s work with distant reading (Cecire 281). Cecire engages with Tanya Clement’s approach to *The Making of Americans*, where Clement visualizes the text in order to take a step away from traditional readings of Stein. Cecire aims to bring in “a history of media, woman’s labor, and the status as literature” as an extension
of this work. Rather than interpreting “unreadability” as a problem of scale, she focuses on the problem of style situated in a history of “compromised reading” in the early twentieth-century (Cecire 283): “And the alternate explanation of Stein’s unreadability holds that she is not impossible to read, but rather ‘difficult.’ One must be willing to put in the work” (301).

Clement and Cecire’s existing conversation on Stein’s readability laid the groundwork for this study. How should we read Stein? And what do nontraditional readings of Stein have to offer? Clement and Cecire suggest that scholars have misread Stein. I agree with Clement that alternative reading practices enable a scholar to track patterns and catch details otherwise undetected by traditional reading practices. Cecire notes the difficulty in reading Stein, but does not explore the kind of work that goes into reading a difficult text. Digital approaches, such as reading with NLTK, offer a way for scholars to grapple with difficult texts.

The Difficulty with Reading Stein

To expand on what Cecire calls a difficult text we can look to Leonard Diepeveen’s definition of “difficulty.” He articulates what difficult means for modernist writers:

*The Difficulties of Modernism,* from its definitions on up, stays close to this sense of difficulty as an experience. It defines difficulty in terms of how modern readers understood and used it: as a barrier to what one normally expected to receive from a text, such as its logical meaning, its emotional expression, or its pleasure. For
modern readers, difficulty was the experience of having one’s desires for comprehension blocked, an experience provoked by a wide variety of works of art (“comprehension” is here defined broadly). Without dealing with this barrier in some way—and such dealings were not restricted to understanding or decoding the syntax of the difficult moment—it was impossible to interact significantly with the text. Difficulty thus drove its readers forward, for they realized that their bafflement was an inadequate response (DM, Diepeveen x-xi).

The “difficult” modernist texts prevent NLP and human readers from comprehending content in different ways. This “barrier” to understanding is an experience recreated when using machine reading with Stein’s writing. This difficulty is what makes errors arise in the process of tagging Stein’s text—above the normal error rate. This difficulty is also what contributes to the nonreading and unreadability Cecire describes of Stein’s texts. Diepeveen studies the cultural importance difficulty played in modernism, and Cecire makes motions to do the same through her argument of labor’s role in the writing and reading of Stein.

The difficulty of reading Stein has been part of a longer conversation. Astrid Lorange notes Stein’s difficulty writing, “[Stein’s] vocabulary is basic—misleadingly so, for she writes complexly and innovatively. Stein’s texts endure. Or perhaps, more exactly, Stein’s readers endure her texts—for they do, undoubtedly, require a significant effort” (Lorange 7). Stein’s “basic vocabulary” demands effort from readers to interpret. In Tender Buttons Stein writes: “The change has come. There is no search. But there is, there is that hope and that interpretation and sometime, surely any is unwelcome, sometime there is breath and there will be a sinecure and charming very charming is that
clean and cleansing” (T n. pg.). Reading Stein is not a sinecure and the task of interpretation is laborious. The basic vocabulary is deceiving and its organization resists interpretation. Punctuation also controls the rhythm of the sentences. Stein shifts from short, concise sentences to long sentences. Lorange talks about how this use of punctuation affects the reader:

Stein had high expectations of her readers, and this in part explains her ambivalence toward punctuation. … Stein’s ideal reader is one who comes to a text willing to wager and err on its tonal, rhythmic, and semantic nuances. Her model reader takes the initiative to pause, breathe, emphasize, speed up, and slow down without explicit instruction (Lorange 97).

While Stein’s ambivalence toward punctuation may signal that readers should find rhythm on his/her own, the punctuation serve as line breaks that we see in work like “Sacred Emily.” Each line in “Sacred Emily” ends with a period before the line break. Her earlier work in “Sacred Emily” signals toward a use of punctuation as more explicit instructions than Lorange suggests. Punctuation is important to account for in critically reading Stein’s writing.

Using NLTK to read Stein’s writing means taking a step away from the original text and critically reading the tagged text. The tagged text serves as a separate layer to Stein’s work rendered visible by NLTK. This difficulty in reading Stein is described by Cecire similar to chores that go unrewarded:

The history of not reading Stein is a history of recognizing the potential unfavorability of the economic exchange on offer, your time and attention for Stein’s writing, all guarantees withheld. We know in advance that the reading will
be repetitive, lengthy, and thin on narrative progression, much like housework and potentially as unrewarded—so great so great, sew grate sew grate. What is so unsettling is how resolutely Stein seems to resist defending her writerly legitimacy in the terms of deliberate, waged or wage-eligible work (301).

Cecire looks to give value to the labor behind repetitive work. Rather than use digital tools, Cecire critiques Clement’s approach and uses traditional reading approaches to study the repetition in Stein’s work. Clement finds this repetition in distant reading *The Making of Americans* and determines that the repetition has a pattern. Cecire is less concerned with the pattern of repetition but rather the repetition itself and the significance of the repetition in reading Stein’s writing.

In “A Poetics of Difference: *The Making of Americans* and Unreadable Subjects,” Melanie Taylor finds Stein’s portraits to be unreadable because the narrator’s gender is ambiguous. Taylor finds repetition is dependent on the “sameness and difference of absolute terms” (Taylor 30). Repetition in Stein’s later prose work is discussed in Stein’s essay “Portraits and Repetition.” Unreadable and nonreading are defined as separate practices that have had an influence over scholarship on Stein. After reviewing the tagged parts of speech from the NLTK analysis, I’m particularly interested in Cecire’s contention that “projected senses of failed reading and writing are genuinely important features of Stein’s texts” (293). I’m interested in how the errors of the tagging function in NLTK and Stein’s style of writing enter into Cecire’s conversation of unreadability.

Stein’s writing breaks rules so errors resist correctness. Her writing doesn’t follow traditional grammar rules and this undoes any mistakes: “To try and discern a mistake does not make sense when standards of linguistic correctness—the very grounds on
which error, as such, is predicated—no longer exist” (Dworkin 116). Her writing becomes legible under these circumstances. Stein writes: “…words began to be for the first time more important than the sentence structure or the paragraph” (P 17). The shuffle of words and “errors” that arise take priority over sentences and paragraphs.

Critical of Clement’s digital methodologies used to analyze The Making of Americans, Cecire acknowledges the significant role human interpretation in machine reading (Cecire 299). Cecire finds these approaches of “nonreading” challenge reading practices in a similar fashion that Stein has challenged reading practices for the last century: “The relationship between not reading and the unreadable, then, is not simply a matter of logic (unreadable texts should not be read), but rather embedded in an early twentieth-century history of compromised reading, whether performed by machines or by women” (Cecire 283). While it is worthwhile to understand how these varying reading practices influence literary scholarship and readings of Stein, not enough attention has been given to the implementation of digital tools used for machine reading and the implications of errors when run with Stein’s texts.

Julie Goodspeed-Chadwick addresses how current scholarship on Tender Buttons is situated between two camps: one that focuses on her use of language and another that focuses on gender relations (Goodspeed-Chadwick 19). She looks to where these two areas of scholarship intersect in Stein’s writing as a site of identity politics at play that “reconfigure[s] language in order to reconfigure representations of identities” (Goodspeed-Chadwick 11). Like Cecire’s work, Goodspeed-Chadwick’s attention to these two camps pieces together a reading of Tender Buttons that signals toward a combination of gender and queer theory with digital approaches.
Parts of Speech in Stein’s Work

Stein foregrounds her concern with parts of speech in her essay “Poetry & Grammar” through classifying “interesting” and “uninteresting” elements of grammar:

A noun has been the name of something for such a very long time. That is the reason that slang exists to change the nouns which have been names for so long. I say again. Verbs and adverbs and articles and conjunctions and prepositions are lively because they all do something and as long as anything does something it keeps alive (LIA 214).

Her essay grapples with how parts of speech function as elements in poetry and prose. The “lively” attribute of these parts of speech works in opposition to the noun. Nouns operate in opposition to the lively parts of speech. The noun is a source of tension, part of a “vocabulary” the writer is charged with to infuse a text (LIA 231). Stein complicates the form of poetry and prose:

And what is the vocabulary of which poetry absolutely is. It is a vocabulary entirely based on the noun as prose is essentially and determinately and vigorously not based on the noun. Poetry is concerned with using with abusing, with losing with wanting, with denying with avoiding with adoring with replacing the noun. … When I said. A rose is a rose is a rose is a rose. And then later made that into a ring I made poetry and what did I do I caressed and completely caressed and addressed a noun (231).

The use and abuse of nouns turns them into “interesting” elements to experiment with. To caress and address the noun becomes an opportunity for error. Stein’s writing, categorized with “radical new aesthetic movements” of the early twentieth-century,
actively reworks complacency of natural language (“Newspaper,” Diepeveen 207). The writing disrupts and distorts both close reading, a decelerated, analytical practice, as well as machine reading, an approach designed by the scholar to address a particular element of a text.

This approach with NLTK developed alongside an attempt to utilize machine reading to address specific aspects of Stein’s works, particularly grammar. Stein’s concern with grammar surfaces in her essay “Poetry & Grammar.” NLTK aligns as a complementary approach to gather large amounts of this information. From the quantitative data, this information was developed from both the pos_tag function in NLTK, as well as tables from R Studio with the imported data:²

<table>
<thead>
<tr>
<th></th>
<th>Verbs</th>
<th>Nouns</th>
<th>Adjective</th>
<th>Determiner</th>
<th>Conjunction</th>
<th>Preposition</th>
<th>Adverbs</th>
<th>Pronouns</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tender Buttons</td>
<td>18.20%</td>
<td>26.97%</td>
<td>7.45%</td>
<td>17.01%</td>
<td>5.34%</td>
<td>7.96%</td>
<td>9.30%</td>
<td>3.68%</td>
</tr>
<tr>
<td>“Sacred Emily”</td>
<td>13.55%</td>
<td>48.67%</td>
<td>6.13%</td>
<td>8.86%</td>
<td>3.33%</td>
<td>4.92%</td>
<td>4.92%</td>
<td>4.92%</td>
</tr>
<tr>
<td>The Autobiography of Alice B. Toklas</td>
<td>18.84%</td>
<td>23.86%</td>
<td>5.99%</td>
<td>10.17%</td>
<td>5.12%</td>
<td>10.79%</td>
<td>7.62%</td>
<td>11.76%</td>
</tr>
</tbody>
</table>

After charting the numbers of parts of speech tagged, there was a twenty-one to twenty-² Outline of methodology and process: Project Gutenberg copies of Tender Buttons, “Sacred Emily,” and The Autobiography of Alice B. Toklas were used with the part of speech tagging function in the Natural Language Processing with Python program. This information was then imported to R Studio to chart the frequency of each tagged part of speech. See Appendix II, III, and IV for the complete list of the part of speech tags for each text used. Punctuation, indentation, and spaces were removed from the results to focus on the tokenized units with words. The percentages for this table are based on this equation: Total number of tokens categorized with a part of speech tag divided by the total number of tokens in the corpus, times one hundred. Data available in csv files upon request.
four percent difference in Stein’s use of nouns between her poetry in “Sacred Emily” and what has been considered one of her more normative works, *The Autobiography of Alice B. Toklas*. Overall, Stein’s usage of nouns has the highest percentage among all parts of speech in the three works analyzed. Stein’s use of pronouns and prepositions increase in *The Autobiography of Alice B. Toklas*. The greatest change in part of speech usage within *Tender Buttons* is in the amount of determiners. Compared to “Sacred Emily,” there is an approximate eight percent increase of determiners in *Tender Buttons*. When compared to *The Autobiography of Alice B. Toklas*, there is an approximate seven percent increase in the use of determiners in *Tender Buttons*. The prose poems that compose *Tender Buttons* have a large amount of determiners interspersed (such as a, an, the). *Tender Buttons* is the exemplary text of the “interesting” use of parts of speech. Lorange outlines what Stein says about parts of speech in Stein’s “Poetry & Grammar” essay. This summary highlights what Stein found interesting and uninteresting in parts of speech:

> It can be summarized by the following: nouns and adjectives are not ordinarily interesting, because they claim to name and describe things directly; verbs are interesting because they are often indirect and so easily mistaken; adverbs, too, are interesting for their capacity to confuse (since adverbs qualify by modulating or transforming rather than simply clarifying or explaining); articles and prepositions are the most interesting of all because they are the least specific, the most context-dependent, and thus most easily mistaken; and punctuation (in particular, exclamation points, question marks, and commas) is generally uninteresting, because it tells a reader when to emphasize, inflect, and breathe—all of which ought to be done without guidance (Lorange 95-96).
Each part of speech has a role in meaning making. Stein is concerned with distancing herself from clarity of meaning and content, and creating “portraits” of objects that are unstable.

Stein’s disinterest in punctuation is also seen in her essays. She writes long sentences, with no commas, that push readers to understand meaning on their own: “And so for me the problem of poetry was and it began with Tender Buttons to constantly realize the thing anything so that I could recreate that thing. I struggled I struggled desperately with the recreation and the avoidance of nouns as nouns and yet poetry being poetry nouns are nouns” (LIA 238). Stein finds nouns inescapable in poetry. Stein’s approach to grammar is best exemplified in the prose poems in Tender Buttons. Nouns are transformed by the abundance of articles and adverbs.

“And so in Tender Buttons and then on and on I struggled with the ridding myself of nouns, I knew nouns must go in poetry as they had gone in prose if anything that is everything was to go on meaning something. And so I went on with this exceeding struggle of knowing really knowing what a thing was really knowing it knowing anything I was seeing anything I was feeling so that its name could be something, by its name coming to be a thing in itself as it was but would not be anything just and only as a name (LIA 242).

The nouns that “must go in poetry” defamiliarize. Here “go” can either mean to get rid of or put in. Stein struggles with whether to keep or rid her work of nouns. Stein concluded
that while nouns are necessary to create meaning, there is a danger of nouns becoming meaningless with overuse. There is a significant decrease in nouns from “Sacred Emily.” *Tender Buttons* has 26.97% nouns compared to 48.67% in “Sacred Emily.” In “Sacred Emily,” the form is more traditional with line breaks. Stein decreased her use of nouns in *Tender Buttons* to prevent the overuse and make the language more forceful through more careful selection of nouns.

The following three charts visualize the percentage of the parts of speech in “Sacred Emily,” *Tender Buttons*, and *The Autobiography of Alice B. Toklas.*

---

**Chart 1. “Sacred Emily” Tagged Data based on Appendix IV.**

---

3 See Appendix I for chart with POS-tagged abbreviations.
In Chart 1 we see that “Sacred Emily” had a word count total of 1,321. There were a total of 21.42% words tagged with NN (noun, singular) and 20.66% tagged as NNP (proper noun, singular). There are 8.86% DT (determiners). In Chart 2 we see Tender Buttons had a word count total of 14,961. 17.01% of words were tagged DT (determiners). Surprisingly, this is nearly twice the amount of determiners that were tagged in “Sacred Emily.” “Sacred Emily” had 20.66% of words tagged as NNP (proper noun, singular) compared to 3.54% of words in Tender Buttons. In Chart 3 we see The Autobiography of Alice B. Toklas Tagged Data based on Appendix III.
Alice B. Toklas had a word count total of 92,835. There are 10.17% determiners and 23.86% nouns. I attribute these differences to the style of Stein’s poetry compared to her prose poems and prose. Her experimentation with parts of speech can be seen through her use of nouns and determiners.

Reading Stein with NLP Data

From the formulation of this machine reading approach, the following section synthesizes the data with a close reading of some of Stein’s work. Articles manipulate nouns in the process of reworking meaning, and determiners are abundant in Tender Buttons. The first prose poem in the “Objects” section of Tender Buttons consists of one header and three sentences. Below each line is the NLTK tagged part of speech:

```
The [DT] difference [NN] is [VBN] spreading [VBG] (T n. pag.).
```

The NLTK program identified the “A” in “A kind in glass and a cousin” as a singular proper noun [NNP], whereas the seven other instances “a” that appear in this section are tagged as a determiner [DT]. NLTK read an instance of a determiner as a noun. The article meant to “please as the name that follows cannot please” is marked with the part of speech it modifies (LIA 212). Stein’s use of grammar and syntax disrupts the
program’s encoded conceptions of natural language. Two types of distortions occur—one by the reader and the other by NLP. The distortion for human readers is intentional. These errors that arise from NLP reinforce the difficulty of reading Stein in both traditional and nontraditional approaches. The arrangement of the system is “not ordinary” yet “not unordered,” alluding to the strangeness of an ordered system. The vessel that is the carafe is “a blind glass” that obscures visibility through non-normative syntax.

At the conclusion of the “Food” section of Tender Buttons, a feminine figure becomes an object to be read in “A Centre in a Table.” Homophony passes between sentences, where “reed” and “read” are juxtaposed and run through repetition:

Suppose a cod liver a cod liver is an oil, suppose a cod liver oil is tunny, suppose a cod liver oil tunny is pressed suppose a cod liver oil tunny pressed is china and secret with a bestow a bestow reed, a reed to be a reed to be, in a reed to be. Next to me next to a folder, next to a folder some waiter, next to a foldersome waiter and re letter and read her. Read her with her for less (T n. pag.).

The text appears to demand for the woman, a feminine figure, to be read. In the succession of phrases within this section, a linkage between nouns grows to “cod liver oil tunny.” This pattern with nouns repeats to form the phrase “foldersome waiter and re letter.” Below is an excerpt of this passage demonstrating how NLTK tagged “reed” and “read”:

\[
\text{[...] secret} \text{ with } a \text{ bestow } \text{ a } \text{ bestow } \text{ reed } \text{ reed } \text{ to to be } \text{ to be, in a reed to be. Next to me next to a folder, next to a folder some waiter, next to a foldersome waiter and re letter and read her. Read her with her for less (T n. pag.).} \]

\[
\text{secret} \text{ with } \text{ bestow } \text{ bestow } \text{ reed } \text{ reed } \text{ to to be } \text{ to be, in a reed to be. Next to me next to a folder, next to a folder some waiter, next to a foldersome waiter and re letter and read her. Read her with her for less (T n. pag.)}. \]

37
and _CC_ letter _NN_ and _CC_ read _VBP_ her _PRP$_0. Read _NNP_ her _PRP$_0 with _IN_ her _PRP$_0 for _IN_ less _IN_ (T n. pag.).

Here “reed” cycles through both verb and noun usage. Ambiguity renders the language susceptible to both human and machine misreadings. The “her” in this passage is read “next to a folder some waiter.” The waiter is “some waiter,” a figure with no clear identity that performs laborious work, and “her” position “next to” the waiter subjects her to this act of being read. While the monosyllabic “re” echoes the sounds from both “reed” and “read,” the e sound in “read” complicates the phrase “read her” as the long e and short e sounds become interchangeable between the confines of the consonants. NLTK reads these homophonic words as both verbs and nouns. The homophony increases their “likelihood of being mistaken” (Lorange 95). On Stein’s use of language and the errors it renders, Astrid Lorange writes:

> She directly associated interestingness with the likelihood of being mistaken:
> language is interesting if it is likely to be misrecognized, read as an error, or read in error; inversely, language is uninteresting when it attempts the direct transmission of unambiguous fact. For Stein, the most interesting words are the words most easily mistaken, and the most interesting composition is the one that favors error-prone words. Her keenness for mistakes is syllogistic: she likes mistakes because they’re interesting, and she is interested in the interesting because interesting things cause errors (95).

Stein composes the prose poems and organizes them to be “misrecognized” by transforming familiar material into unconventional meanings. Where the phrase, “...a reed to be. Next to me...” shifts sounds between “be” and “me,” new patterns emerge
with the “foldsome waiter.” Whether the speaker proposes to read “her” or already has read “her,” misrecognition occurs on a phonetic level that doesn’t register with NLTK while performing these functions. In the “Rooms” section of *Tender Buttons*, the speaker contemplates why “the name” changes: “Why is the name changed The name is changed because in the little space there is a tree, in some space there are no trees, in every space there is a hint of more, all this causes the decision” (T n. pag.). As a noun, “the name” stands in for “the name of something for such a very long time” (*LIA* 214). The noun takes a generic form, and the speaker struggles to grapple with an arbitrariness that dictates how “the name” changes over time. Homophony and repetition disrupt how NLTK reads Stein’s works. Repetition is weaved throughout the prose poems, and becomes part of the syllogistic keenness of Stein’s for “error-prone words” (Lorange 95). Stein’s arrangement of the parts of speech renders the text susceptible to misrecognition in machine reading.

*Tender Buttons* was one of Stein’s earlier works. She experiments with language to create more “interesting” writing that is error-prone. From the poem “Sacred Emily” there is a change in these prose poems. In “A Transatlantic Interview” her efforts in *Tender Buttons* was noted:

*The early book, Tender Buttons, was written in Spain in 1913 and was Gertrude Stein’s first attempt to “express the rhythm of the visible world.”* Tender Buttons was, therefore, to Gertrude Stein’s development what the “Demoiselles d’Avignon” was to Picasso’s, a key work marked with the enormous struggle of creating a new value (P 23).
Part of this “new value” in her writing comes from the work described in Lectures in America. Through tagging the texts we can see a progression in how her use of parts of speech and language shifts over genre and time. This change can be seen in charts 1-3 above. The form of her writing changes and Stein uses longer sentences instead of the short sentences seen in “Sacred Emily.”

In “A Transatlantic Interview” from A Primer for the Gradual Understanding of Gertrude Stein, Stein responds to random portions of Tender Buttons. Here is one of the excerpts from Tender Buttons that Stein responds to:

\[
\text{A DT BROWN NNP. } \\
\text{A NNP brown NN which WDT is VBZ not RB liquid JJ not RB more RBR so RB is VBZ } \\
\text{relaxed VBN and CC yet RB there RB is VBZ a DT change NN, a DT news NN is VBZ } \\
\text{pressing VBG (T n. pag.).} ^4
\]

She uses several adverbs in this passage above—a part of speech Stein finds interesting. The adverbs transform the noun “brown.” The two uses of the “not” adverb are in opposition of each other negating meaning. The verbs “relaxed” and “pressing” are also in opposition. In response to this prose poem Stein says: “The color is held within and there you see I was groping for the color” (P 27). Like the first prose poem in “Objects,” the NLTK program identified the “A” in “A brown which is not liquid…” as a singular proper noun [NNP], whereas the two other instances “a” that appear in this section are tagged as a determiner [DT]. This shifts the meaning—from “A DT BROWN NNP” that denotes a common brown to “A NNP brown NN” that alludes to a specific brown.

---

^4 Tender Buttons is quoted in the interview. I have cited from Tender Buttons and added the POS-tag from NLTK.
Below is another passage that Stein responds to in the interview as an attempt at “sound pictures”:

PEELED\textsubscript{VBD} PENCIL\textsubscript{NNP}, CHOKE\textsubscript{NNP}.

Rub\textsubscript{NNP} her\textsubscript{PRPS} coke\textsubscript{NN} (T n. pag.).

The last four single syllable words are sharp. Of the six total words, four are nouns, one is a pronoun, and one is a verb. Stein says she gives up on this style: “That is where I was beginning and went on a gool [good] deal after that period to make sound pictures but I gave that up as uninteresting” (P 27). The “sound picture” is heavy with nouns and has little “interesting” parts of speech. “PEELED” is the single verb NLTK identified. “Rub” is identified as a noun rather than a verb. “CHOKE,” too, is identified as a noun. If this is read as a section with the majority of nouns, this passage offers a clear collage of images. The previous prose poem analyzed, “A brown which is not liquid…,” is less like one of these “sound pictures.”

At the end of “Objects” Stein uses commas to punctuate the sound in her prose poem:

\textsc{THIS, IS, THE, DRESS, AIDER}.

Aider\textsubscript{NNP}, why\textsubscript{WRB} aider\textsubscript{NN} why\textsubscript{WRB} whow\textsubscript{NN}, whow\textsubscript{NN} stop\textsubscript{NN} touch\textsubscript{NN}, aider\textsubscript{NN} whow\textsubscript{NN}, aider\textsubscript{NN} stop\textsubscript{NN} the\textsubscript{DT} muncher\textsubscript{NN}, muncher\textsubscript{NN} munchers\textsubscript{NNS}.

A\textsubscript{DT} jack\textsubscript{NN} in\textsubscript{IN} kill\textsubscript{NN} her\textsubscript{PRPS}, a\textsubscript{DT} jack\textsubscript{NN} in\textsubscript{IN}, makes\textsubscript{NNS} a\textsubscript{DT} meadowed\textsubscript{VBN} king\textsubscript{VBG}, makes\textsubscript{NNS} a\textsubscript{DT} to\textsubscript{TO} let\textsubscript{VB} (T n. pag.).
Verbs are mistaken for plural nouns. Stein’s writing leads the NLTK POS-tagger to incorrectly categorize parts of speech. Her concern with the “use and abuse” of nouns leaves ambivalence to their use and category.

With the NLP data we learn more about Stein’s writing style through the shift in her noun usage between “Sacred Emily,” *Tender Buttons*, and *The Autobiography of Alice B. Toklas*. We learn that there is an increased error rate in NLP with Stein’s texts. The meaning of passages changes when read with NLP.

Homophony and repetition also contribute to readings of Stein as composition stemming from women’s labor. It is important to understand this aspect from the mechanization aspect underlying Stein’s work. Cecire focuses on Stein’s use of homophony in “Sacred Emily” to frame Stein’s poem as one of oikonomia, defining it as “one interested in particular in the household management undertaken by ‘wives of great men’” (286). Cecire gives attention to the opening of “Sacred Emily,” as well as the line “So great so great Emily. / Sew grate sew grate Emily” to highlight the “greatness” derived from “an incantation of repetitive household drudgery” (*GP* 183; Cecire 287).

Repetition is weaved throughout the prose poems, and becomes part of the syllogistic keenness of Stein’s for “error-prone words” (Lorange 95). The arrangement of parts of speech factor into Stein’s preference for misrecognition. Along with “a domestic, largely monosyllabic and bisyllabic lexicon,” Stein’s writing mimics the labor of women, “rendered invisible, female writing was prone to being described as a mere emanation of the body and erased from the literary canon” (Cecire 289, 291). Looking to the repetition in “Sacred Emily,” female writing struggles to be “put” in the hand of the speaker:
Put something down.

Put something down some day.

Put something down some day in.

Put something down some day in my.

In my hand.

In my hand right.

In my hand writing.

Put something down some day in my hand writing (GP 186).

Through each line words are passed down as units keeping the content of these eight lines interconnected. The speaker will “some day” receive agency to write “something,” but only after putting “something down some day.” The speaker carries this “something” that prevents her from writing. Writing is a practice disconnected from her hand. The switch between “right” and “writing” transfers agency away from a purely domestic practice of repetition to a place where “in” the speaker’s hand there is a space for her “hand writing.” Her hand is what writes, and the homophony in Stein’s writing plays on the “invisibility and interchangeability” of a woman’s contribution in feminized occupations (Light 459).

On the history of computation during the early twentieth-century, Jennifer S. Light discusses the consequences of feminized clerical labor as it related to domestic work:

With feminization came a loss of technical status, since other men doing more “important” technical and classified work remained in noncombatant positions. Thus, the meaning of “wartime labor shortage” was circumscribed even as it came into being. While college-educated engineers considered the task of computing
too tedious for themselves, it was not too tedious for the college-educated women who made up the majority of computers (461).

Computing, a task “too tedious” for male labor, is passed to women during this time. The women became “human computer[s]” where the tedious clerical work of physical programming offered “slightly higher status and pay than other kinds of clerical labor” (462). Stein’s automation in her writing engages with the automation of information work. Stein’s reputation of illegible handwriting, itself unreadable “until transmuted by her typists,” calls for attention to distinguish how the unreadable elements of her work by contemporary readers are not due to illegibility (Cecire 299). It is the syllogistic relationship with the interesting and uninteresting parts of speech that motivate Stein to cause misrecognition and errors that render her texts unreadable.

Despite machine reading having some “distance” from the content of a text, this computer-aided reading delves into how Stein’s unconventional writing challenges both traditional and alternative reading practices. Cecire posits that perhaps Clement’s distant reading work is “an experience no human reader is having” (299). She writes, “In the world of the machine, seeing is reading” (300). Thus, comprehension plays a key role for Cecire as she evaluates alternative reading practices when applied to the unreadable. Cecire classifies Clement’s work as a nonreading approach, one that Clement pursued due to misreadings of Stein’s writing in an attempt to “restage” perspective on Stein’s work (299). Machine reading with NLTK not only provides distance from unreadable works, but also allows the reader to control distance and scale. The tagged texts require both human interpretation and close reading in order to piece together quantitative
results. This combination supplements how alternative reading practices factor into larger conversations concerning digital approaches to studying literature.

Stein’s poetics collide with alternative reading practices. Her writing challenges both conventional and nonconventional approaches, and disrupts the “seeing” of machine reading that utilizes Natural Language Processing. Her concern with grammar and vocabulary are rendered visible with computer-aided reading, revealing patterns that reinforce her commentary in “Poetry and Grammar.” Stein’s writing demands special attention in traversing the balance between traditional, close reading and computational approaches.

Machine reading has difficulty with experimental writing. Through the tedious, repetitive labor interlaced with the style of her writing, Stein’s writing serves as a nexus for studying women’s work and machine reading rooted in early twentieth-century American history. The women’s labor, as well as the “daily life” of Americans, differentiates her writing from the English literature inherited by the British. Just as Stein grappled with how developments in writing during the past centuries influenced twentieth-century literature, computation has left an impression on alternative reading practices today. In her essay, “What is English Literature,” Stein contemplates the effect of the nineteenth-century on writing, and writes, “Perhaps we are still under its shadow a little bit” (LIA 41).
Conclusion

With more time and technical resources, a project like this could investigate a writer’s oeuvre. A project like this could also study a larger volume of works. If I had the resources, I would have worked with a linguist and a data scientist to thoroughly check for more patterns and have improved data visualizations. There are stories in data. Often it requires experimentation to extract these stories. I view this project as a first step as part of a larger investigation.

As a graduate student in an English Master’s program, a great deal of labor went toward learning how to program and use Natural Language Processing for the purpose of this study. I did not have a strong computer science background. I believe with additional support more patterns in the existing data could be revealed. Collaboration with others is necessary for stronger interdisciplinary work. Natural Language Processing is one of many digital tools that can be used in digital approaches to literary studies.
## Appendix I

<table>
<thead>
<tr>
<th>Tagged Abbreviation</th>
<th>Part of Speech</th>
</tr>
</thead>
<tbody>
<tr>
<td>CC</td>
<td>coordinating conjunction</td>
</tr>
<tr>
<td>CD</td>
<td>cardinal number</td>
</tr>
<tr>
<td>DT</td>
<td>determiner</td>
</tr>
<tr>
<td>EX</td>
<td>existential there</td>
</tr>
<tr>
<td>FW</td>
<td>foreign word</td>
</tr>
<tr>
<td>IN</td>
<td>preposition/subordinating conjunction</td>
</tr>
<tr>
<td>JJ</td>
<td>adjective</td>
</tr>
<tr>
<td>JJR</td>
<td>adjective, comparative</td>
</tr>
<tr>
<td>JJS</td>
<td>adjective, superlative</td>
</tr>
<tr>
<td>LS</td>
<td>list marker</td>
</tr>
<tr>
<td>MD</td>
<td>modal</td>
</tr>
<tr>
<td>NN</td>
<td>noun, singular or mass</td>
</tr>
<tr>
<td>NNS</td>
<td>noun plural</td>
</tr>
<tr>
<td>NNP</td>
<td>proper noun, singular</td>
</tr>
</tbody>
</table>

---

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>NNPS</td>
<td>proper noun, plural</td>
</tr>
<tr>
<td>PDT</td>
<td>predeterminer</td>
</tr>
<tr>
<td>POS</td>
<td>possessive ending</td>
</tr>
<tr>
<td>PRP</td>
<td>personal pronoun</td>
</tr>
<tr>
<td>PRP$</td>
<td>possessive pronoun</td>
</tr>
<tr>
<td>RB</td>
<td>adverb</td>
</tr>
<tr>
<td>RBR</td>
<td>adverb, comparative</td>
</tr>
<tr>
<td>RBS</td>
<td>adverb, superlative</td>
</tr>
<tr>
<td>RP</td>
<td>particle</td>
</tr>
<tr>
<td>TO</td>
<td>to</td>
</tr>
<tr>
<td>UH</td>
<td>interjection</td>
</tr>
<tr>
<td>Tagged Abbreviation</td>
<td>Part of Speech</td>
</tr>
<tr>
<td>--------------------</td>
<td>--------------------------------</td>
</tr>
<tr>
<td>VB</td>
<td>verb, base form</td>
</tr>
<tr>
<td>VBD</td>
<td>verb, past tense</td>
</tr>
<tr>
<td>VBG</td>
<td>verb, gerund/present participle</td>
</tr>
<tr>
<td>VBN</td>
<td>verb, past participle</td>
</tr>
<tr>
<td>VBP</td>
<td>verb, sing. present, non-3d</td>
</tr>
<tr>
<td>VBZ</td>
<td>verb, 3rd person sing. present</td>
</tr>
<tr>
<td>WDT</td>
<td>wh-determiner</td>
</tr>
<tr>
<td>WP</td>
<td>wh-pronoun</td>
</tr>
<tr>
<td>WP$</td>
<td>possessive wh-pronoun</td>
</tr>
<tr>
<td>WRB</td>
<td>wh-adverb</td>
</tr>
</tbody>
</table>
Appendix II

*Tender Buttons*, List of Part of Speech Tags

<table>
<thead>
<tr>
<th>POS_tag</th>
<th>Number of Tokens Tagged</th>
</tr>
</thead>
<tbody>
<tr>
<td>CC</td>
<td>799</td>
</tr>
<tr>
<td>CD</td>
<td>71</td>
</tr>
<tr>
<td>DT</td>
<td>2411</td>
</tr>
<tr>
<td>EX</td>
<td>240</td>
</tr>
<tr>
<td>IN</td>
<td>1191</td>
</tr>
<tr>
<td>JJ</td>
<td>947</td>
</tr>
<tr>
<td>JJR</td>
<td>130</td>
</tr>
<tr>
<td>JJS</td>
<td>37</td>
</tr>
<tr>
<td>MD</td>
<td>66</td>
</tr>
<tr>
<td>NN</td>
<td>3170</td>
</tr>
<tr>
<td>NNP</td>
<td>530</td>
</tr>
<tr>
<td>NNPS</td>
<td>1</td>
</tr>
<tr>
<td>NNS</td>
<td>335</td>
</tr>
<tr>
<td>PDT</td>
<td>4</td>
</tr>
<tr>
<td>POS</td>
<td>2</td>
</tr>
<tr>
<td>PRP</td>
<td>493</td>
</tr>
<tr>
<td>PRPS</td>
<td>11</td>
</tr>
<tr>
<td>RB</td>
<td>1130</td>
</tr>
<tr>
<td>RBR</td>
<td>85</td>
</tr>
</tbody>
</table>
The total number of tokens in the text (excluding punctuation and spaces). → **14961**
Appendix III

*The Autobiography of Alice B. Toklas*, List of Part of Speech Tags

<table>
<thead>
<tr>
<th>POS_tag</th>
<th>Number of Tokens Tagged</th>
</tr>
</thead>
<tbody>
<tr>
<td>CC</td>
<td>4751</td>
</tr>
<tr>
<td>CD</td>
<td>831</td>
</tr>
<tr>
<td>DT</td>
<td>9045</td>
</tr>
<tr>
<td>EX</td>
<td>389</td>
</tr>
<tr>
<td>FW</td>
<td>2</td>
</tr>
<tr>
<td>IN</td>
<td>10021</td>
</tr>
<tr>
<td>JJ</td>
<td>5276</td>
</tr>
<tr>
<td>JJR</td>
<td>163</td>
</tr>
<tr>
<td>JJS</td>
<td>129</td>
</tr>
<tr>
<td>MD</td>
<td>820</td>
</tr>
<tr>
<td>NN</td>
<td>10940</td>
</tr>
<tr>
<td>NNP</td>
<td>7873</td>
</tr>
<tr>
<td>NNPS</td>
<td>91</td>
</tr>
<tr>
<td>NNS</td>
<td>3250</td>
</tr>
<tr>
<td>PDT</td>
<td>70</td>
</tr>
<tr>
<td>POS</td>
<td>352</td>
</tr>
<tr>
<td>PRP</td>
<td>8764</td>
</tr>
<tr>
<td>PRP$</td>
<td>1582</td>
</tr>
<tr>
<td>RB</td>
<td>6339</td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td>----</td>
<td>----</td>
</tr>
<tr>
<td>RBR</td>
<td>112</td>
</tr>
<tr>
<td>RBS</td>
<td>31</td>
</tr>
<tr>
<td>RP</td>
<td>361</td>
</tr>
<tr>
<td>TO</td>
<td>2614</td>
</tr>
<tr>
<td>VB</td>
<td>2753</td>
</tr>
<tr>
<td>VBD</td>
<td>8682</td>
</tr>
<tr>
<td>VBG</td>
<td>1240</td>
</tr>
<tr>
<td>VBN</td>
<td>2435</td>
</tr>
<tr>
<td>VBP</td>
<td>1197</td>
</tr>
<tr>
<td>VBZ</td>
<td>1184</td>
</tr>
<tr>
<td>WDT</td>
<td>369</td>
</tr>
<tr>
<td>WP</td>
<td>556</td>
</tr>
<tr>
<td>WPS</td>
<td>17</td>
</tr>
<tr>
<td>WRB</td>
<td>596</td>
</tr>
</tbody>
</table>

The total number of tokens in the text (excluding punctuation and spaces). -> 92835
Appendix IV

“Sacred Emily,” List of Part of Speech Tags

<table>
<thead>
<tr>
<th>POS_tag</th>
<th>Number of Tokens Tagged</th>
</tr>
</thead>
<tbody>
<tr>
<td>CC</td>
<td>44</td>
</tr>
<tr>
<td>CD</td>
<td>3</td>
</tr>
<tr>
<td>DT</td>
<td>115</td>
</tr>
<tr>
<td>EX</td>
<td>1</td>
</tr>
<tr>
<td>IN</td>
<td>65</td>
</tr>
<tr>
<td>JJ</td>
<td>71</td>
</tr>
<tr>
<td>JJR</td>
<td>7</td>
</tr>
<tr>
<td>JJS</td>
<td>3</td>
</tr>
<tr>
<td>MD</td>
<td>11</td>
</tr>
<tr>
<td>NN</td>
<td>283</td>
</tr>
<tr>
<td>NNP</td>
<td>273</td>
</tr>
<tr>
<td>NNS</td>
<td>87</td>
</tr>
<tr>
<td>PRP</td>
<td>52</td>
</tr>
<tr>
<td>PRP$</td>
<td>6</td>
</tr>
<tr>
<td>RB</td>
<td>55</td>
</tr>
<tr>
<td>RBR</td>
<td>3</td>
</tr>
<tr>
<td>RBS</td>
<td>1</td>
</tr>
<tr>
<td>TO</td>
<td>47</td>
</tr>
<tr>
<td>VB</td>
<td>48</td>
</tr>
<tr>
<td>Tag</td>
<td>Count</td>
</tr>
<tr>
<td>-------</td>
<td>-------</td>
</tr>
<tr>
<td>VBD</td>
<td>19</td>
</tr>
<tr>
<td>VBG</td>
<td>19</td>
</tr>
<tr>
<td>VBN</td>
<td>4</td>
</tr>
<tr>
<td>VBP</td>
<td>33</td>
</tr>
<tr>
<td>VBZ</td>
<td>56</td>
</tr>
<tr>
<td>WDT</td>
<td>2</td>
</tr>
<tr>
<td>WP</td>
<td>7</td>
</tr>
<tr>
<td>WRB</td>
<td>6</td>
</tr>
</tbody>
</table>

The total number of tokens in the text (excluding punctuation and spaces). -> **1321**
Bibliography


