EXPLORING SENSES, EMOTIONS, AND THEIR INTERCONNECTIONS THROUGHOUT LITERARY PERIODS

BY

AKHILA ASHOKAN

THESIS

Submitted in partial fulfillment of the requirements
for the degree of Master of Science in Computer Science
in the Graduate College of the
University of Illinois Urbana-Champaign, 2022

Urbana, Illinois

Adviser:

Professor Roxana Girju

ABSTRACT

Literature presents a unique platform to experience senses and emotions. Particularly, fiction literature brings alive sensory perception and emotional experiences in its language. This study reports on the semantic organization of English sensory descriptors of the five basic senses, their interconnections, and their associations with emotions in a large corpus of over 7,000 Project Gutenberg fiction books. In addition, this study reports on a distributional-semantic word embeddings approach to identify and extract these descriptors and analyze their mixing interconnections in the resulting conceptual and sensory space in three important periods: 1700s, 1800s, and 1900s, as well as overall. In the analysis, I attempt to understand the how the semantic sensory space is organized for each literary period and the general sensory interactions seen across all literary periods. This work also tries to align the sensory spaces of different literary periods with literary movements. Furthermore, I explore different techniques to extract and rank the sensory descriptors seen in each corpus and describe the strengths and drawbacks of each approach. This research is novel in the the visualizations, which catch a glimpse of the perceptual spaces of sensory experiences in fiction narratives. The results are presented in multiple formats, including insightful and interactive interfaces for colorful data visualization and linguistic analysis. The findings are relevant for research on concept acquisition and representation as well as for applications that can benefit from a better understanding of perceptual spaces of sensory and emotion experience in fiction, in particular, and in language, in general.

To my family for their support and encouragement during this journey.
Thank you.

ACKNOWLEDGMENTS

Many thanks to my advisor Dr. Roxana Girju. This thesis would not have been possible without her guidance, support, and insight. I appreciate her taking the time to read and review this thesis. She has held up the proverbial torch by guiding me on this journey. I would also like to thank the University of Illinois Department of Computer Science and Graduate College for their continued assistance and patience in this process. Finally, I would like to thank my family, friends, and loved ones for giving me the strength and love to complete this journey.

TABLE OF CONTENTS

CHAPTER 1: INTRODUCTION

Every interaction, regardless of how small, is perceived in the human mind through a sensory receptor. Our ability to perceive is so deeply intertwined with our senses that a loss of one sense can result in the strengthening of others [1]. Sensory experiences are translated through language into writing across many cultures and languages, making written text a primary source of sensory language. A skillful writer's use of sensory language can captivate a reader, making them feel, understand, and reason with the tone of text. This leads me to believe sensory language is the medium through which humans communicate experiences, evoke emotions, and build empathy. It is also possible that our use of sensory languages has changed with the tides of human evolution, and consequently, taken new forms based on the contemporary social environment. Our senses are critical to the human experience, yet, there is limited work mapping sensory spaces in texts across different time spans. Studies have suggested that an empirical study mapping the sensory space could enlighten us on the techniques used to bring forth emotions and provide clues on how to strengthen empathy in writing [2]. In this thesis, I present my research on the coverage of the five basic senses in fictional literature, spanning the 1700s, 1800s, and 1900s. Preliminary findings of this study were presented in [3].

1.1 WHY FICTION?

Written texts are abundant in sensory language, but there is a case to be made for fiction literature, in particular. Well-written fiction literature causes the reader to become completely immersed in a fabricated reality. Whether it be the ballrooms of eightieth-century London or the space stations of futuristic Mars, readers are drawn into these worlds and held captive for pages. Creating the kind of universe that one is entirely unfamiliar with requires the use of descriptions. Any imagery created with these descriptors appeals to a basic sensory modality or sometimes several. It's not possible to describe a location that one has never been to or experienced without explaining the awe-inspiring sights or sensational sounds. Fiction's goal is to push the mind's limits to new extremes, and it accomplishes this using dense sensory imagery [4].

Another reason why fiction was selected for this study is that it can be used to understand the ideas of a particular time. In fact, fiction embodies the social and intellectual ideas of a time in it's settings, characters, and plot. For example, it can show contemporary attitudes towards government or the role of women in society [5]. In addition, fiction, like many works of literature, is influenced by literary movements. This study looks at each literary period's sensory space to determine if the influence of literary movements can be seen in the sensory language.

1.2 THE FIVE SENSES FOLK MODEL

This study will map the sensory space using the traditional five senses folk model as described by Bodo Winter [6]. These fives senses, sight, touch, hearing, smell, and taste, are also popularly known as the "Aristotelian" senses. This organization of the five senses is based on individual sensory organs. I choose to use this model for simplicity during analysis, but it should be noted that adopting this perspective has its own flaws. For one, the five sense model is not held up in all cultures [7]. This does not pose an significant issue for this study since the data set is in English. There has also been a scientific studies indicating more than one neuropsychological area may be responsible for sensory stimuli [6]. For example, pain, which is thought to be related to touch in the five senses model, has been linked to structures outside those associated with the general sensation of touch. Cross-modal interactions further complicates this classification system. A perfect classification system for the senses does not exist, but it's important to note the complexity of this space. For the purposes of this study, a five sense system helps to analyze the text data in a clear manner.

1.3 SYNESTHESIA IN LITERATURE

As mentioned previously, sensory experiences in the body do not naturally fall into the five basic sense categories. In language, we see a similar phenomenon in synesthesia. Synesthesia is a literary device that creates cross-sensory metaphors using two or more sensory modalities [8]. Neurological synesthesia has been used to describe individuals who experience synesthetic perceptions in daily life. The topic of neurological synesthesia continues to be a subject of debate and is studied by the psychologists, neuroscientists, and philosophers [9]. In contrast, literary synesthesia has been remarked by some as "intersense analogy" that requires the ability to differentiate between senses in a way that contradicts the symptoms of neurological synesthesia [10].

Literature contains an abundance of synesthesia. The American poet, Emily Dickinson, was known to include literary synesthesia in her work [11]. For example, Dickinson writes, "I clutched at sounds", imposing tactile words on an auditory sensation - an unusual combination of senses. Dante Alighieri's 'The Divine Comedy' is an earlier work containing a plethora of literary synesthesia [2]. However,synesthesia makes a prominent appearance in literature during the ninetieth century [11]. At this time, the industrial revolution and the rise of the scientific discovery pressed artists to think outside the confines of reality and experiment with their imagination. J.-K. Huysmans' 1884 novel, *Against Nature* exemplifies this in one quote among many others, "Indeed, each and every liquor, in his opinion, corresponded in taste with the sound of a particular instrument...crème de menthe and anisette like the flute, at once sweet and tart, soft and shrill. . . " [11, 12]. Synesthesia has also appeared in more recent thrillers like T.Jefferson Parker's *The Fallen* [11]. In this book, Detective Robbie Brownlaw's investigation is described using synesthetic langauge:"The red squares of the lie spilled from his mouth" [13].

Despite the prevalence of synesthesia as a literary device, the underlying mechanisms of this device are not well studied. This study examines synesthesia from a broader perspective, referred to as sensory blending from here on out. Sensory blending can be understood as a the various interconnections between individual senses. It can include metaphors and similes like those seen in synesthetic language, but also refers to senses which appear in the same contexts. This study will map out the sensory blends that occur in texts from the 1700s, 1800s, and 1900s using a data-driven approach.

1.4 RESEARCH CONTRIBUTIONS

This study was designed with four research questions in mind. RQ1 deals with understanding the sensory modalities in general. In order to answer RQ1, the data set is split into literary periods. Within the interactive figures for each period, all five sensory modalities are mapped and their interactions can be clearly delineated. The results illustrate the overarching patterns of sensory blending and which modalities are solitary, if any. RQ2 compares each literary period to discern similarities among them and to highlight the differences. RQ3 investigates whether literary movements impact the sensory language. If language is a dynamic entity that changes with the existing cultural climate, there is a chance that sensory language may also be influenced. Finally, RQ4 examines how effective each of the ranking methods applied are in identifying the top descriptors of a sensory modality.

RQ1 How is the semantic sensory space organized in each literary period?

RQ2 What are the general sensory interactions seen across all literary periods?

RQ3 How do literary movements influence the sensory descriptors seen in a period?

RQ4 How well do the approaches utilized in this study capture the sensory descriptors?

CHAPTER 2: LITERATURE REVIEW

This chapter will focus on related works in the fields of computational linguistics, natural language processing (NLP), and literature. First, I will discuss works that have previously explored sensory language. This section will be followed by a overview on literary movements broadly and a section on NLP techniques for literature. Finally, I will end with a discussion on the applications of this work.

2.1 SENSORY LANGUAGE IN TEXT

Sensory language is studied by linguistics and experts in literature, but only a few major works delve into understanding this area by applying computational approaches. One work provides the basis for my study and looks at English sensory words in the same data set as the one used in this study [2]. In this work, Girju et al. applies a word2vec approach to map out the organization of sensory words without distinguishing between literary periods. Other studies survey the sensory space beyond the five senses folk model. In addition to the basic sensory modalities, English perception verbs can be classified as agentive and experiential. One study examines these English perception words using a word embedding approach to uncover the similarities and differences between these perception verbs, further highlighting the complexity of sensory language [14]. Other studies apply computational approaches to understand one sense in particular: smell. Rather than focusing on all five basic senses, a study by Brate et al. examines the presence of olfactory language in Project Gutenberg novels [15]. This paper contributes a data set of annotated references to smell and shows how iterative bootstrapping techniques can be used to identify olfactory language. Olfactory language has also been studied in the context of psychophysical properties by Iatropoulos et al. [16]. This study has supplied a validated metric to characterize olfactory related words using a psychophysical data set. In fact, contributions from this study are extended in [14] and applied as one of the ranking methods I apply in my study (PAI). While this study provided a data-driven method for identifying odor descriptors, it does not provide additional information on how these descriptors are semantically organized. In a related study, odor descriptors are automatically identified using the metrics presented by Iatropoulos et al. and are semantically organized based on their distribution in natural English texts [17]. Work published as recently as 2021 has built upon these prior contributions and the sights of domain specialists to build annotation guidelines for olfactory texts [18]. These guidelines were then applied to a set of travel writings to analyze the olfactory experiences in texts.

Table 2.1: A General Classification of Literary Movements in the 1700s, 1800s, and 1900s

Literary Movements	**English Literary Movements**	**American Literary Movements**
1700	Neoclassical	Puritan or Colonial
1800	Romanticism, Realism, Naturalism, Victorian	Transcendentalism, Realism, Naturalism
1900	Modernism, Postmodernism	Modernism, Postmodernism

This study sets up a framework for future work to automatically label texts in a supervised manner, a task that has not been successfully tackled on sensory language.

Despite the limited number of computational studies in this space, sensory linguistics has been studied for a long period of time and there are a vast number of resources on language and perception in general. This is simply because humans have perceived the world through their senses for as long as they have been around. One such resource was created by Bodo Winter and describes what should be studied under the discipline of ”sensory linguistics” and how it should be studied. While part one of the book is largely theoretical, part two provides a detailed account of English sensory terms and their properties [6].

2.2 LITERARY MOVEMENTS

A key piece of my study is comparing and contrasting between different literary periods to see what distinctions appear. Literary periods are better presented in the context of literary movements for the purposes of this study since literature is strongly influenced by the historical context in which it was written. Although I look at each period as one distinct literary period, each of these periods is actually composed of several different literary movements happening in America and England. It’s important to note that literary critics disagree on the time periods associated with each movement and the exact defining characteristics. Literary movements do not fall strictly within the centuries investigated and the convention of periodical classification can be problematic. Overlapping movements and timelines make this analysis difficult, but it is still possible to categorize these literary periods in terms of common themes or literary trends that existed at that time. One of the main research questions I seek to answer is whether specific sensory descriptors or patterns can be attributed to a literary movement. At the same time, I wish to see what patterns characterize all the literary periods and the human experience of perception as a whole. To strengthen the foundation for my discussion on literary periods, I describe some key writings and studies on literary periods and the characteristics of each periods explored in this study.

Most literary movements in English can be broken into periods of English literature and periods of American literature. Author Mario Klarer describes each of these eras in his book introducing literary periods to a novice reader [19]. Movements in English literature can be listed in order starting from the eighteenth century as: neoclassical, romantic, Victorian, realism, naturalism, modernism, and postmodernism. The major periods of American literature in approximate chronological order are: colonial or Puritan, romantic, transcendentalism, realism, naturalism, modernism, and postmodernism. The neoclassical age of literature, also called eighteenth century, golden, or Augustan age literature, applies themes from a classical literary era to contemporary society. While neoclassical literature continued in England, Puritan or colonial age literature developed in America following the rise of religious ideology in the area. These early American texts provide a clearer understanding on the role theology played in non-European society. Romanticism arose after the eighteenth century in England and is marked by ”nature and individual, emotional experience” [19]. In short order, transcendentalism developed in America in the first half of the nineteenth century. Perhaps influenced by Romanticism, this area of work explains philosophy through the lens of nature. Following this, England and the Americas followed the same literary trajectory: realism, naturalism, modernism, and postmodernism. Realism, as the term suggest, tries to authentically portray reality through language. In contrast, naturalism depicts how social and environmental factors influence characters and in turn, human life [20]. While realism and naturalism arose, the Victorian age in England formed in the later half of the nineteenth century, combining aspects of realism and modernism to show life under the reign of Queen Victoria. The transition into the twentieth century was accompanied by the rise of modernism and new writing style techniques like literary cubism and stream of consciousness writing. The same novel writing techniques introduced in the era of modernism were formalized and revived in the postmodern era.

While I have attempted to summarize these eras, various surveys may choose to describe these eras differently. There are limited empirical studies to classify texts into literary periods. In one study, Amancio et al. classifies books published from 1590 to 1922 into six clusters of books that coincide with well-established literary movements seen in the last several centuries [21]. The results from this study show these five literary movements spanning the 1700s to the 1900s: Neoclassicism/Enlightenment (1660–1798), Gothic fiction (1764–1820), Realism (1830–1900), Naturalism (1865–1900), and Modernism (1890–1940).

2.3 NLP TECHNIQUES FOR LITERATURE

In this study, I use existing methods in natural language processing to analyze the sensory space in fictional literature and literary periods. Previous works in computational linguistics have contributed to understanding literature as a whole using similar techniques. The amount of work that has been done at the intersection of literature and computation linguistics is dense so I will describe the notable works that are more directly related to the study conducted here.

The common approaches to NLP tasks are statistical, rule-based, traditional machine learning based, or deep learning based. One study has applied an unsupervised approach (narrative modeling) which utilizes Latent Dirichlet Analysis to disentangle complex narratives in the fictional text*Infinite Jest* [22]. Another technique that's been introduced into the field is graphs, maps, and trees. Elson et al. reports on a method to build conversation networks using dialogue interactions in literary fiction [23]. Elson et al.'s findings were valuable because they reveal that a majority of literary texts do not fit the descriptions ascribed to them by literary experts using an empirical method. Extrapolating on this idea, my study aims to see if literary hypotheses, presented in regards to sensory information and literary movements, can be validated by examining the sensory spaces of literary period. Another work that extends on Elson et al.'s study investigates how computational techniques can be applied to validate literary theories [24]. However in this work, the authors expand the set of literary theories and delve deeper into the creation of social networks from literary texts. They experiment with interaction and observation networks and apply a new network extraction method that they find to be more effective at validating literary theories.

Other studies have used lexical and syntactic clues within the text as features in a supervised approach to identify the speakers in a novel [25]. Supervised learning models have also been used to classify literary texts by perceived literariness using textual features [26]. The models used to classify these texts were not complex and simple bi-grams and support vector machines could be used. Furthermore, these results could also be obtained for content and stylistic features of texts. Beyond text classification, pre-trained transformer based models, like BERT and ELECTRA, can assist in single-label emotion classification for distinctive texts like plays. These findings suggest that a supervised approach can be applied to explore literature [27, 28].

Software tools have also been built to assist in the adoption of NLP tools for literary researchers. GutenTag is a standalone software for non-programmers that allows researchers to easily interact with the texts in Project Gutenberg [29]. As mentioned later on in this paper, Project Gutenberg is a valuable resource in literary studies because of the vast number of plain-text manuscripts available for public use. GutenTag's interface allows users to create a corpus that fits their analytic needs with little too no coding required. This tool is aimed at bridging the gap between the digital humanities and computation linguists as mentioned by Hammond et al [30]. Despite the contributions from both communities, Hammond et al. suggests that there is still a barrier between the communities that does not allow literary scholars to adopt and utilize the tools produced by computational linguists.

2.4 APPLICATIONS

In this last section, I explain the different applications of this work on sensory spaces in literature. In particular, I focus on three main areas where this work can be applied in the long-term. However, other areas may benefit from an understanding of how humans use sensory language.

2.4.1 Adaptive Technologies

As humans become more involved in a world of digital technology, they become more reliant on these technologies to communicate. Thus, it becomes a necessity for tools to meet the needs of all those who are using these digital platforms. Tools need to able to process human emotions and experience in order to meet these needs. As stated in previous works, human emotions have a direct link to human sensory experience [2]. Feelings are often explained through literal and figurative sensory language. At the same time, sensory language can prompt emotional responses. Studies in education, medicine, and extended reality already call for the integration of multimodal interactons that appeal to all the senses [31, 32, 33]. Furthermore, adaptive and multimodal technolgoies have the potential to assist those with visual and auditory impairments by compensating for their sensory void. Cognitive linguists have noted that augmenting a blind child's sensory experience with related sensory experiences has the potential to improve their comprehension of reality [34]. This study serves as a step-stone for creating such adaptive technologies, particularly text-based technologies, that appeal to all human senses and emotions.

2.4.2 Emotional and Sensory Systems

In exploring how humans use sensory language, we hope that this study and others like it can provide insight into how emotional and sensory systems work. Despite how ingrained sensory experience is in daily life, there is little to tell us about how the body experiences senses and it's link to emotion. Furthermore, the study of emotion is still a largely open area of research. Questions still remain about how sensory systems work together to elicit an emotional response in our minds. Some have suggested that emotions themselves are embedded in our evolutionary history [35]. If this is the case, examining how we display sensory language over the years can give clues on how much evolution has shaped human emotion and it's expression in sensory language.

2.4.3 Creative Writing

The use of sensory language can also immerse readers in the scenes that authors envision. A good writer has a creative use of language that evokes emotions and even empathy in readers. This study presents a good starting point for understanding how writers are able to blend together senses in their writing. It also serves to develop creative writing tools. Computational techniques to approach creative writing have always been viewed with some skepticism. However, introducing these new tools for writing can provide perspective to new writers and help them develop strategies for good writing [36]. Writing, of course, can be a hobby or an therapeutic endeavor for the writer. It is evident from prior research that depicting sensory details with creativity can have a beneficial effect on trauma survivors by creating a easy platform for self-expression [37]. Multimodal sensory dairies have also been found to improve the well being of teachers who utilize them to navigate and process emotional stress [38].

CHAPTER 3: APPROACH

My approach to understanding sensory perception across literary periods focuses on the five basic sensory modalities. I applied a semi-automatic method to extract sensory descriptors from a corpus of fiction texts and applied five different ranking methods to reduce noise in the data set and filter out irrelevant descriptors. Then, I used a word embedding model to map the sensory space and principal component analysis (PCA) to visualize the interactions among sensory descriptors across periods of time. In the next sections, I explain the data collection process, the ranking techniques, the computation model, and the visualization steps.

Table 3.1: Number of Texts In Each Literary Period

Literary Period	Number of Texts
1700	314
1800	6475
1900	698
Entire Corpus	7487

3.1 DATA

For this study, I utilized the Gutenberg, dammit corpus[1], which consists of plain text Project Gutenberg ebooks collected until June 2016. The corpus comes packaged into multiple sub-directories and the individual manuscripts are identified with unique, numerical Gutenberg IDs. The corpus also provides official Project Gutenberg metadata on each text in a JSON file. The entire corpus contains 50,729 manuscripts written by 18,462 unique authors. I narrowed down the corpus to include only books with ‘Fiction’ or ‘fiction’ in the genre list and with ‘English’ as the only language within the text. If multiple genres were included, I selected texts with at least one fiction genre. The metadata did not provide information on publication date so I determined the approximate publication century by examining the author’s birth year. It was required that a valid author birth be given to at least one author so that each manuscript could be assigned to a literary century. If a text had multiple authors born in different centuries, I used the first listed author’s birth date to determine the text’s publication century. Texts that were written prior to the 1500s were binned together. The majority of texts fell in the 1700s, 1800s, and 1900s bins so I selected

[1] The Gutenberg, dammit (https://github.com/aparrish/gutenberg-dammit) was created by Allison Parrish and licensed under the Creative Commons Attribution-ShareAlike 4.0 International License (https://creativecommons.org/licenses/by-sa/4.0/)

Table 3.2: Top 20 Genres

Genre	Frequency
Fiction	804
Science fiction	715
Short stories	670
Adventure stories	295
Love stories	251
Historical fiction	233
Detective and mystery stories	230
Conduct of life – Juvenile fiction	221
England – Fiction	188
Western stories	164
Man-woman relationships – Fiction	156
England – Social life and customs – Fiction	139
Domestic fiction	135
Humorous stories	135
Young women – Fiction	109
English fiction – 19th century	108
England – Social life and customs – 19th century – Fiction	106
Children – Conduct of life – Juvenile fiction	106
Psychological fiction	103
London (England) – Fiction	94

these three centuries for further analysis. Table 3.1 shows the distribution of the corpus based on the author's birth century. With these requirements, 7,487 manuscripts written by 1,956 unique authors were selected for further analysis. The top 20 authors and top 20 genres in the filtered corpus are shown in Tables 3.3 and 3.2 respectively. The filtered corpus, with the labeled author birth century, was tokenized and part of speech tagged using the en_core_web_sm pipeline provided in the spacy library [39]. In order to compare and contrast the literary periods, I separate the final corpus by the three major author birth centuries (i.e. 1700, 1800, 1900) into three separate corpus, one for each literary period. For the reminder of the steps in our pipeline, I perform the same set of steps on each corpus in addition to the the combined corpus containing all the English fiction texts.

3.2 SEED WORDS

In order to extract descriptors relevant to the five sensory modalities, I applied a semi-automatic approach to select one set of non-overlapping seed words that are semantically and morphologically linked to each of the five sensory modalities. I started with a set of seed words manually selected for each sensory modality: sight (see/look), hear (hear/listen/sound), touch (touch/feel), taste (taste/flavor/savor), smell (smell/scent/odor). Then, I branched out from these base seed words to semantically related words using the WordNet lexical database. The WordNet lexical database is freely accessible as part of the nltk library [40]. I used the WordNet database to automatically identify related words like hyponyms and hypernyms. Phrases that were linked to the base seed word were removed. I utilized Python's word form library to automatically find morphological variations of the words. The final seed words list included 873 seed words across all modalities with 150-200 words per modality. A sample set of seed words can be seen in Table 3.4. Each of the seed words was then part of speech tagged using the en_core_web_sm pipeline provided in the spacy library [39]. The same set of seed words was used for all four corpora.

Table 3.3: Top 20 Authors

Author	Frequency
Georg Ebers	125
George Meredith	100
W. W. (William Wymark) Jacobs	90
Mark Twain	87
George Manville Fenn	65
Anthony Trollope	64
G. A. (George Alfred) Henty	62
Charles Dickens	61
Henry James	50
William Le Queux	50
Gilbert Parker	50
Edward Stratemeyer	47
Charles James Lever	47
Bret Harte	46
Eugène Sue	42
Henry Rider Haggard	41
Randall Garrett	38
Winston Churchill	37
Robert Louis Stevenson	35
Roy J. (Roy Judson) Snell	35

Table 3.4: Sample Seed Words

Sense	Count	Sample Seed Words
sight	190	see,look,witness,glance,stare,glimpse,admire,visualize,peer,behold
hear	174	hear,listen,noises,music,strum,twang,ring,voice,tone,gong
touch	192	touch, brush, press, palpate, feel, stroke, hit, kiss, strike, fingertips
taste	163	taste, seasoning, sweet, sour, salty, vanilla, bitter, lemony, savor, mellow
smell	154	smell, odor, aroma, sniff, scent, nose, acrid, incense, rancid, perfume

Table 3.5: Number of Descriptors Extracted From Each Literary Period

Literary Period	**Min Count Threshold**	**Sight**	**Hear**	**Touch**	**Taste**	**Smell**
1700	2	6242	5063	3613	1602	679
1800	30	6167	3976	3269	1054	759
1900	2	781	422	607	121	134
Entire Corpus	30	6452	4239	3460	1149	794

3.3 EXTRACTING DESCRIPTORS AND CONTEXT WINDOWS

To extract sensory descriptors from each corpus, I programmatically extracted content words that surround a seed word. Specifically, I looked at a context window of +- 4 centered around a seed word. Context windows were truncated if a sentence boundary was encountered and descriptors were only selected if they occurred above a set threshold. The threshold was set based on the size of the corpus. The smaller corpora from the 1700s and 1900s were given a threshold of 2. The larger corpora from the 1800s and the full corpus was given a threshold of 30. The threshold was used to reduce the computation complexity and improve the feasibility of further calculations, but the threshold also helped to prune words that occurred less frequently within sensory context windows. In addition, descriptors words were only selected if part of speech tagged as either a noun, verb, adjective, or adverb. The total number of sensory descriptors extracted from each period after excluding descriptors outside the threshold is shown in Table 3.5. The seed words used to center context windows were not considered part of the context window, but they were considered descriptors if they occurred within context windows. The window size was selected based on the parameters set for this data set in previous works, but the parameters should be further tested and validated for other data sets [2]. Each period and modality differed in the size of the set of extracted context windows. Notably, the 1800s literary period and the full corpus provided the largest context window sets. The smallest context window set came from the 1900s literary period, despite the 1900s data set being the second smallest data set. The log-scaled distribution of sensory imbalance is shown in Figure 3.1.

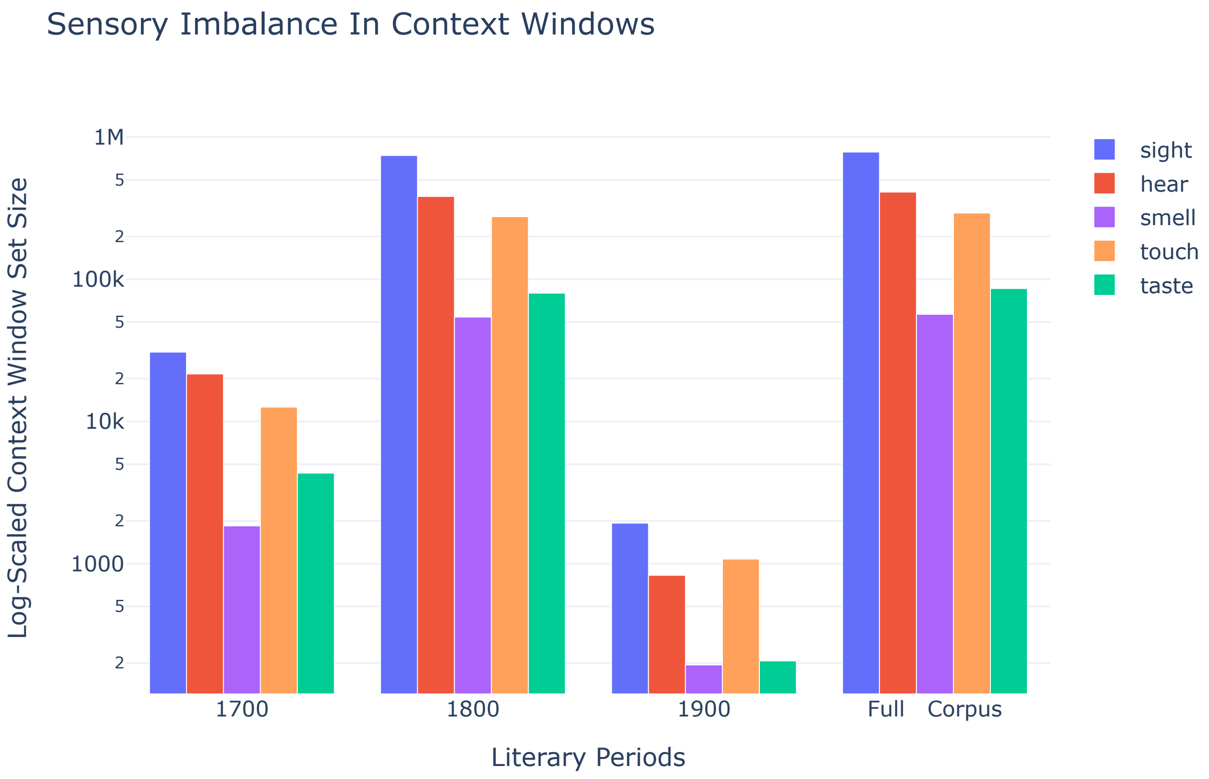

Figure 3.1: Log-scaled size of context window set extracted from each corpus. In general, we extracted the largest context window set from sight seed words. The 1900s produced the smallest set of context windows.

3.4 DESCRIPTOR RANKING METHODS

To visualize the top sensory descriptors in each period, four ranking methods were applied. After plotting the PCA for the top descriptors, I examined plots to determine whether they represent the sensory modalities. In this section, I describe each method in-depth and how it was implemented.

3.4.1 Perception Association Index(PAI)

Prior works have discussed how informative sensory descriptors can be identified to provide a list of top descriptors [2, 14, 41]. In the first ranking method, I ranked the list of descriptors by calculating the Perception Association Index (PAI) for each descriptor and selecting the top sensory descriptors with the highest PAI value. The term Perception Association Index was originally developed and used in a previous study on sensory spaces in the English language [14]. PAI is an extension the Olfactory Association Index (OAI), introduced in a study conducted by Iatropoulos and colleagues on the semantic content of olfactory words [16]. OAI measures how strongly a word is associated with the idea of smell. Note that this metric was validated on psychophysical datasets and literature has shown that high OAI ideas are linked to high rankings of olfactory association. PAI, which applies OAI to the other four senses as well, is the log2 probability that a descriptor d occurs in a sensory context rather than a non-sensory context. PAI allowed me to select a descriptors with a higher correlation to sensory semantics and visual them more easily on PCA. PAI is calculated as follows, where d is the descriptor, $p_f(d)$ is the frequency of the descriptor in sensory contexts, and $t_f(d)$ is the total frequency of the descriptor.

$$PAI(t) = log2(\frac{p_f(d)}{t_f(d)}) \tag{3.1}$$

3.4.2 TF-IDF

I also used another method to rank the top descriptors in each literary period, TF-IDF. Traditionally in natural language processing tasks, TF-IDF is used to rank documents, words, and phrases in a corpus of documents. The TF in TF-IDF stands for term frequency. TF measures the relative frequency of a word in a document. There are many variations of TF. TF can be represented as the raw count of a term in a document, the count of a term in a document normalized by document length, or as a logarithmally scaled frequency among other variations. The IDF in TF-IDF is the inverse document frequency. IDF weighting is added to the formula to allow rare but informative words to take precedence over common and uninformative words. The standard definition of IDF is simply the number of documents in the corpus, N, divided by the document frequency $df(t)$ or the number of documents in which the term, t, appears.

$$IDF(t) = \frac{N}{df(t)} \tag{3.2}$$

By applying TF-IDF in different ways, I developed four different TF-IDF-based ranking methods and tested out each on the data set. It was determined that the documents in each ranking method would be the context windows extracted for each sensory modality. For example, the document for sight in the 1700s literary period would include all the sight context windows extracted from the 1700s corpus. Following this, the entire corpus for the 1700s literary period would be the collection of five context window sets. A seperate document six was also added to some ranking methods. This document six was created from any sentences in the entire corpus that did not contain a sensory seed word. This essentially allowed us to consider all the non-relevant, non-sensory terms as a separate document.

Now, I describe the four TF-IDF ranking methods. The first two methods are the simple and were implemented without using any external libraries. Method one utilized the five documents created using the sensory context windows. It is defined as the the term frequency multiplied with the inverse document frequency with an additive smoothing factor applied. The additive smoothing factor is applied to the denominator of the IDF to avoid a zero division error for terms that don't appear in the document set. In method two, a similar implementation is used but the sixth document of non-relevant sensory descriptors is added to the document set. The TF-IDF implementation used in method one and method two is shown below.

$$TF - IDF(t, D) = tf(t, d) * idf(t, D) = \frac{c(w, d)}{|d|} * \frac{N}{df(t, D)} \tag{3.3}$$

Methods three and four are implemented using the scikit-learn TfidfVectorizer. The smoothing parameter was enabled and the l2 euclidean norm was utilized to normalize for the different document lengths. The difference between methods three and four is that method four also includes the non-sensory context windows in document six as part of the document set.

3.5 COMPUTATIONAL MODEL

Using to my advantage the presence of sensory specific context words around chosen seed words, I identified descriptors that express sensory content. I was then able to map out their semantic organization by training word embedding models on the extracted context windows. The word embedding model represents each word as a separate vector distance in a multi-dimensional space. This means words with a similar semantics will have similar vector representations. I trained a word2vec with CBOW model on contexts windows using a hidden layer with 200 units, minimum word count of 1, and 30 training iterations. The model was trained on the set of all context windows extracted from each corpus with no distinction between senses. The distance between two descriptors, i and j, was calculated using p, the Pearson correlation between word vectors, as shown below:

$$D_{ij} = 0.5 * (1 - p_{ij}) \tag{3.4}$$

Following this, the distance D between descriptors is converted to 0-1 range, with 0 indi-

Table 3.6: Explained Variance Ratio Explained By PCA Models

Literary Period	**Explained Variance Ratio**
1700	0.9568
1800	0.2857
1900	0.74
Entire Corpus	0.2888

cating semantic identity and 1 indicating semantic opposition. The resulting distance matrix is then further visualized using Principal Component Analysis (PCA) with 2-components. In the next section, I briefly explain why PCA was applied before moving on to the results of the PCA in the next chapter.

3.6 PRINCIPAL COMPONENT ANALYSIS (PCA)

In this study, PCA is applied to visualize the top sensory descriptors in each literary period. This technique is applied following previous work conducted by Girju et. al on sensory blending in fiction [2]. Due to it's simplicity, PCA is a common statistical process that has been applied to everything from computer graphics to the COVID-19 virus [42, 43]. My aim in applying it to this study is to understand the underlying relationships between the senses. PCA reduces the dimensionality of the distance matrices that are produced from the computational model. It accomplishes this by condensing the data in the matrix down to 2 vectors that combine all the relevant information in the matrix [44]. I utilized the open-source scikit-learn PCA package [45]. The 3-component PCA was also tested out with the hope that it would separate different sensory patterns. However, the visual results of the 3-component PCA were difficult to analyze and patterns were not easily discernable. Therefore, I opted to use 2-component PCA. To understand how well PCA captures the data that is being visualized, the explained variance ratio was measured for each literary period. Explained variance ratio is the percentage of variance explained by the 2 components [45]. Conveniently, this statistic is available as part of the PCA package provided by scikit-learn. The explained variance ratio for each literary period is reported in Table 3.6. It's important to note that the 1700 and 1900 period PCAs were able to explain a greater portion of the variance than the larger corpora. One possible explanation for this variation is that the 1700s and 1900s corpora are smaller data sets so the 2-component PCA is able to capture more information.

CHAPTER 4: RESULTS AND DISCUSSION

In this chapter, I describe the experimental results from this study. I examined a number of different metrics on the the sensory descriptors extracted from our data set to identify and understand the underlying sensory patterns in fiction texts. These metrics will present us with empirical evidence that answers the initial research questions. First, I present the PCA plots generated from each of the ranking methods discussed in Chapter 3. Then, I will present three new metrics describing the sensory space.

4.1 TOP DESCRIPTORS RANKED WITH PAI

In this section, I will explain the results from using the PAI ranking method using colorful PCA graphs. Using this ranking method, I plotted the top 500 descriptors in each literary period as shown in 4.1. Note that the individual descriptors that make up these plots can be magnified, viewed, and analyzed using the interactive plotting tools used in this study.

The 1700s literary period shows a dense ‘c’ shaped cluster on the left side with five descriptors hanging to the far right. The sparse cluster of descriptors to the right include two body part descriptors: *ear* and *nose*. Moving to the cluster on the left, a heavily overlapping cluster of descriptors appears. Interestingly, sight descriptors appeared to be more spread out compared to the other sensory modalities. Within these dispersed sight descriptors appeared *wistfully, fixedly, glisten, riveted, agape*. As the plot progresses towards the left and towards the origin, we see stronger interactions between the other four senses (hearing, touch, taste, and smell), but less distinctions between specific sensory clusters. Moving on to the 1800s, a clearer set of sensory interactions appear. This PCA plot appears more like a Venn diagram with each modality taking up a part of the plot, but also mixing with each other. It’s important to note that this is the largest corpora that was examined in this study. Starting from the lower left corner of the plot, sight descriptors heavily mingle with touch and hearing descriptors. Many of the sight descriptors seen on the lower left corner seem to represent words that are distinctly related to the eyes (*dilated, narrowed, squinting, stonily, bloodshot*). Progressing upwards in the plot, sight descriptors fall off to be replaced by smell (*acrid, puff, sniffing*) descriptors that are interacting with hearing (*swish, footfalls, rumbling*) and touch descriptors (*index, tips, thumb*). Finally, at the top middle of the plot, there is a clear interaction between smell and taste descriptors. Many of the taste descriptors in this part of the plot are food-related (*sandwiches, berries, and soup*). This literary period showed some clear sensory interactions: 1) sight, hear, touch, 2) smell, touch, hear 3) smell

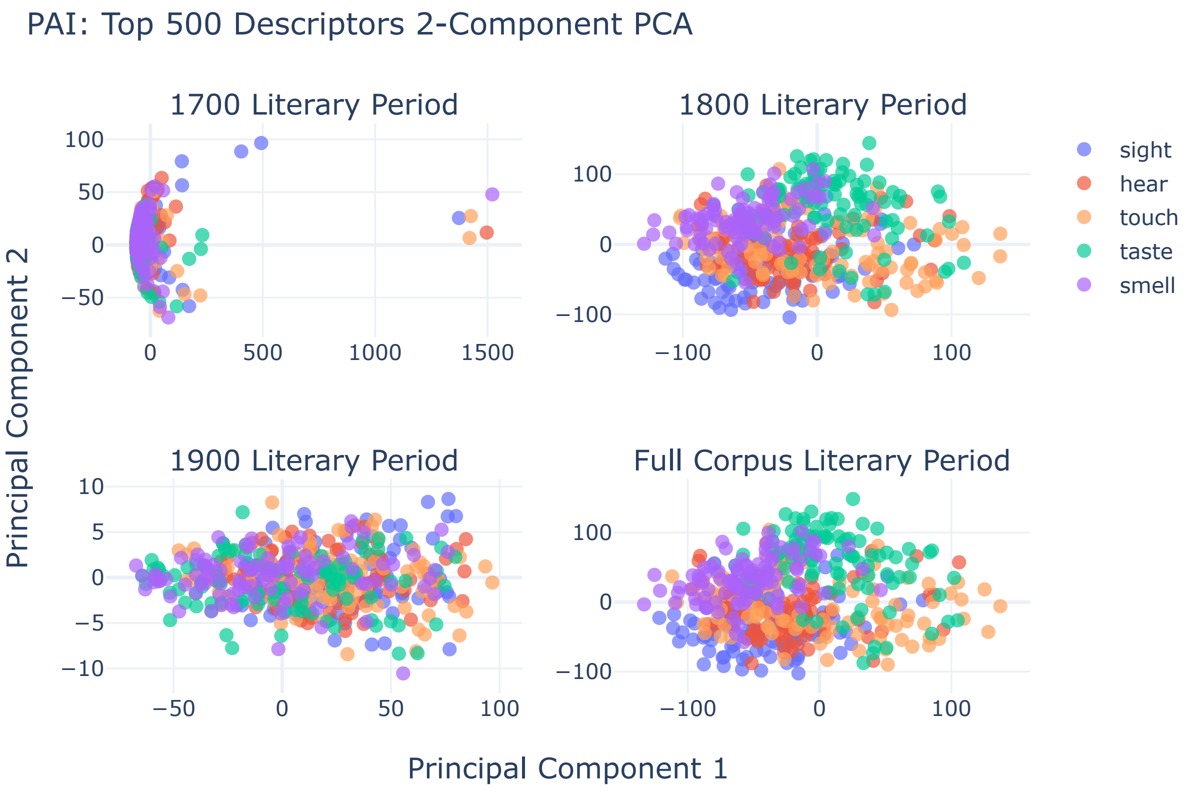

Figure 4.1: Top 500 Descriptors Ranked Using PAI

and taste. The 1900s literary period shows a similar set of patterns. To the left side of the plot, taste and smell descriptors mingle. As we move to the right, smell descriptors mingle with touch and hearing descriptors. Then, the plot grows outward on the y-axis as the sight descriptors come in. Since the full corpus consists most of 1800s texts, the full corpus displays the same patterns that the 1800s plot illustrates.

From examining the results of the PAI plots, I concluded PAI would rank descriptors highly even when they were not very frequent in the corpora. For example, the top 500 descriptors in the 1700s period included *doorways* although it occurred only twice in the corpora. In the this example, the PAI score was one because it occurred in two separate context windows that were extracted from the same sentence. Another reason for this behavior could be that PAI does not take into consideration the term frequency of other descriptors in the corpora. On the other hand, the results from the larger corpora (1800s) show better descriptor rankings and more distinct clusters suggesting that PAI may still be a viable method if the data set is large enough. Now, I will discuss the TF-IDF based ranking methods.

4.2 TOP DESCRIPTORS RANKED WITH TF-IDF

In this study, I experimented with four TF-IDF methods by varying the documents used in the calculations and the underlying normalization. Each method is described in detail in Chapter 3. Methods one and two varied only the inclusion of a sixth document of non-sensory context windows. Similarly, methods three and four only varied in this way. Due to how similar the PCA plots for these methods turned out, I will focus on the results from method two and method four. The PCA plots for methods one and three can be seen in Figure A.1 and Figure A.2.

4.2.1 Method Two

The PCA plot for the 1700's literary period looks similar to the plot shown for PAI. Again, we observe a 'c' shaped curve on the left side of the plot with one smell descriptor, *nose*, on the right hand side. Upon closer examination of the 'c' curve on the left side, it's clear that the sight descriptors are more dispersed. As the plot moves to the left, gradually more hear descriptors join alongside touch, smell, and taste descriptors. Once again, the clusters don't form discernible patterns. I hypothesize that this corpus is does not form the complete picture of the sensory space since it is the smallest of the three periods. The 1800s presents a familiar and interesting picture. Mainly, the same progression of sensory transitions that were observed in the 1800s PAI PCA is seen when moving from the bottom left, clockwise. Again, a similar set of eye-related descriptors (*unflinching, piercing, scowling*) can be seen on the bottom left. It's worth highlighting that these descriptors are also more clearly linked to displays of emotion. This finding is not surprising considering how the the eyes are commonly considered "the window to the soul". Furthermore, psychophysiological studies show that the eyes play a key role in emotional regulation [46]. The interactive plots designed for this study allow the user to see how each sense disperses individually and while alongside the other senses. It's worth noting that sight and touch are the most dispersed. Sight descriptors start with emotional meanings on the left sight and gradually disperse into more figurative language like *feast out eyes, practiced eyes*, etc. The 1900s literary period PCA plot for method two is unlike the PAI plot for the same literary period. In fact, the PAI plot for the 1800s displays the same sensory interactions as the PAI plot the 1900s if in a different orientation.(see Figure 4.1). Figure 4.2 shows, however, that sensory interactions between all five senses exists in the corpus. On the left side, there is a smattering of descriptors from all five senses. Taste descriptors are more prominent towards the right side of the graph. However, these descriptors are more loosely tied to the taste modality (ex: *dress, resumed,*

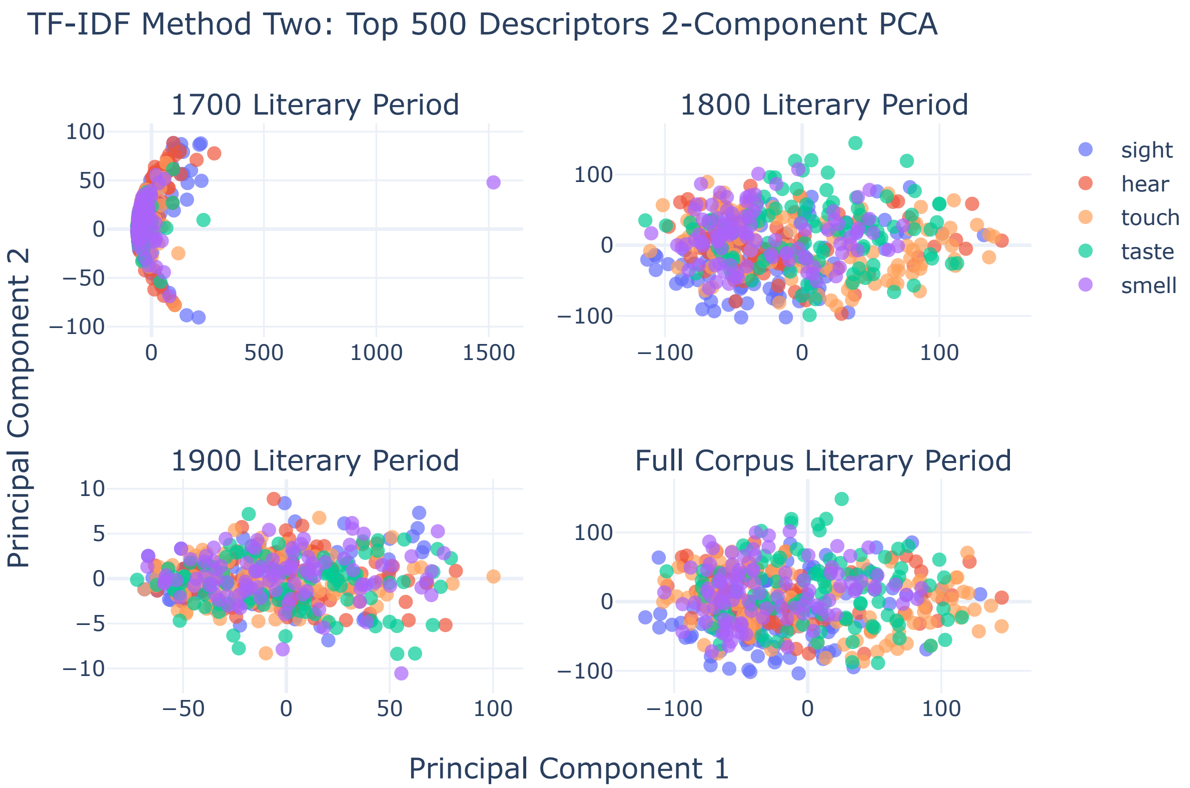

Figure 4.2: Top 500 Descriptors Ranked Using TF-IDF Method Two

triumphed), despite being found in sensory contexts (ex: *dress and eat, resumed eating, bitterness and triumph*). Finally, the PCA for the entire corpus showed similar patterns to that of the 1800s periods.

Analyzing the TF-IDF plots for method two revealed that the sensory blending seen in PAI plots could also be observed using this ranking method. In addition, the types of highly ranked descriptors were the same for the large corpora. However, TF-IDF did a better job of eliminating those descriptors that occurred rarely in context windows. Manual inspection of the interactive PCA plots revealed that more frequent words were captured when compared with PAI. With this, I will move on the analysis of the TF-IDF method four.

4.2.2 Method Four

TF-IDF method four differs from method two in one major way. After computing the TF-IDF score, the scores are normalized by taking the Euclidean norm [45]. Due to this, the PCA plot for the 1700s literary period was more varied than what was seen when using method two or PAI. In the 1700s literary period, a cloud of sensory descriptors appear from mix modalities. Towards the left side of the plot, smell and taste descriptors congregate. These descriptors appear to transition from literal meanings (*fragrant powder, eating cold meat, tastes as like sour milk*) to metaphoric meanings (*face is just as sweet, delicacy of men, biting words*). In particular, the taste descriptors on the left side of the plot seem to be almost entirely food related (*wine, milk, meat, oat*). Whereas, the taste descriptors dispersed out on the right side of the plot are more general descriptors that can take a metaphoric shape (*man, girl, sweet, woman*). The word embedding computation model used for the study is less likely to group these metaphoric descriptors together despite their relevance to the sensory modality. Moving on to the 1800s literary period, we come back to a similar cluster of sensory blending that we've seen in method two and PAI when moving clockwise from the bottom left. Sight descriptors intermingle with touch and hearing descriptors. In this plot, the touch/smell pattern is less apparent as touch mingles evenly throughout the cloud of descriptors. As before, there is less smell and sight interaction. Moving onto the 1900s, the descriptors seem to intermingle uniformly throughout the graph. Examining the interactive plots, it's easier to see the progression of each modality. Sight descriptors cluster heavily towards the left side before dispersing to the right. For example, words like *hand, little, big, and man, face, white* are seen on the left. The touch and hearing descriptors take on a similar dispersion pattern, but they are spread further away from the x-axis. Taste and smell descriptors are more uniformly dispersed without distinct clusters forming on either side of the plot. Again, I examine the full corpus to determine if any unique patterns exists that are not seen in the other corpora. However, this plot mirrors that of the 1800s to a large extent.

Analyzing the PCA plots for TF-IDF method four has shown that applying a normalization step in the ranking process can make a difference in the resulting ranking scores. This effect can be noted in the 1700s literary period more than the others. Furthermore, this affected the resulting sensory interactions in the PCA plots. From examining the plots for methods two and four, it is unclear which method brought out the true sensory patterns in the 1700s, if there is such a thing. However, it is too early to discount that the normalization step helped clarify the sensory patterns seen in a smaller data set like the 1700s literary period.

4.3 SENSE PAIRS

To understand the full picture of sensory interactions, I examine the sensory interactions through the lenses of sense pairs. Specifically, I wanted to examine the extent to which concepts (or descriptors) could be experienced through more than one sensory modality. In

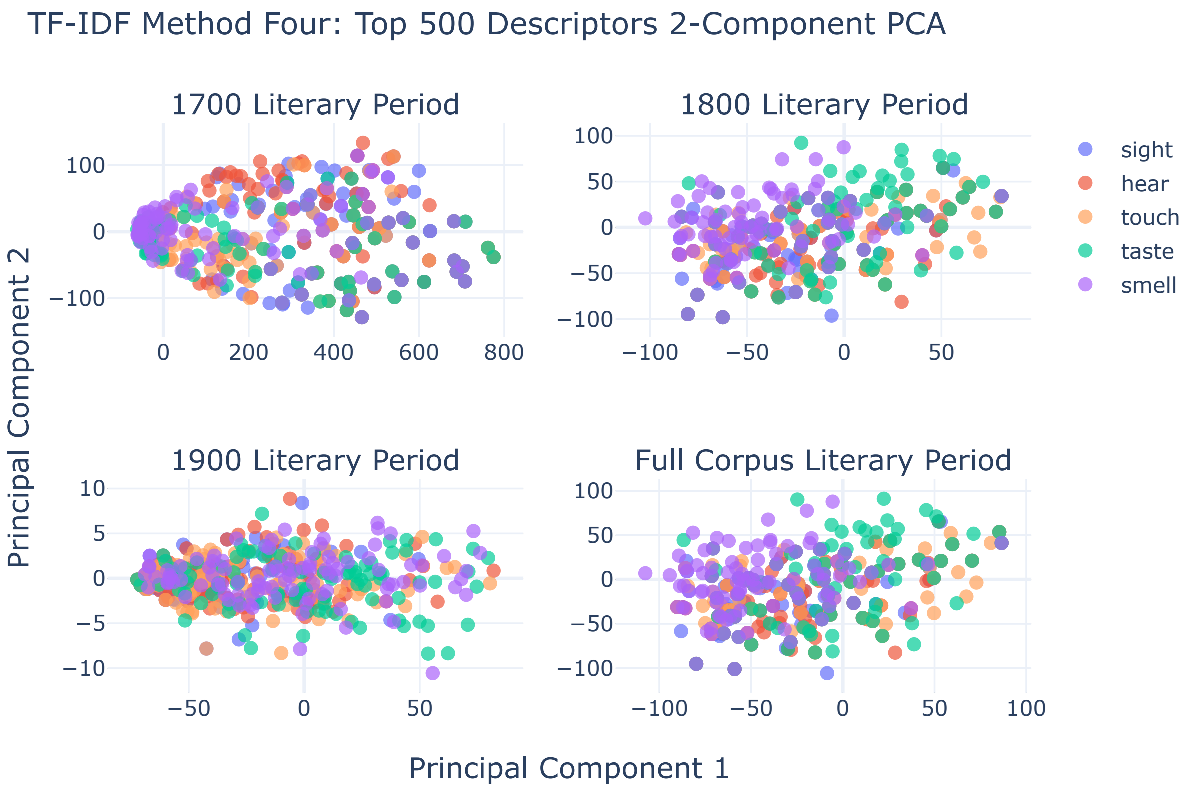

Figure 4.3: Top 500 Descriptors Ranked Using TF-IDF Method Four

fact, studies in neurobiology highlight that sensory modalities do not act independently [47]. Rather, the cross-modal interactions in the brain are the norm. In the sections below, I explore this using two approaches from [2] on the top 500 descriptors ranked using PAI and TF-IDF method four.

4.3.1 Average Distance Between Sense Pairs

The first approach was used to measure the proximity of descriptors to one another including descriptors that belonged to another sense. To achieve this, I calculated the euclidean distance between the top 500 descriptors ranked using PAI and TF-IDF method four and averaged the values for each possible sensory modality pair. This was computed for each literary period.

First, we can observe the results for descriptors ranked with PAI as shown in Figure 4.4. This shows the normalized average distance for different sense pairs in all literary periods. Sense pairs consisting of the same sensory modality are shown in magenta and sense pairs consisting of different modalities are shown in gray. First, we examine the the 1700s literary period. Interestingly, these sense pairs showed very close average distances with some exceptions. The descriptors belonging to the taste-taste group had the lowest average distance. This could indicate that descriptors of taste clustered together more strongly and showed strong semantic similarity. The second smallest sense pair was smell-taste. The smell-taste combination is a strong pattern that can also be observed in the PCA plots. In fact, a strong link between gustation and olfaction has been shown to exist in neuropsychology studies [48]. Sight-sight, sight-touch, and touch-touch sense pairs were found to have the greatest average distance. This is further validated by the scattered dispersion of the sight and touch descriptors in the PCA. In the 1800s, smell-smell, hear-hear, and sight-sight descriptors have the smallest average distance. However, the mixed modal pairs with the smallest distance are hear-sight, hear-smell, hear-touch. These same groupings were observed in the PAI PCA plots. Sight-taste, taste-touch, and hear-taste had the greatest average distance in this period. Moving onto the 1900s, we see a similar group of sensory pairs with the smallest distance (hear-hear, hear-sight, sight-sight). The similarity between the 1800s corpus and the full corpus can be seen in the last graph where the sense pair ordering are almost exactly the same except for the switching of sight-taste and sight-touch.

In order to see if these same sensory pairings can be observed with another ranking method, we look at the normalized average distance for different sense pairs creating using TF-IDF method four as shown in 4.5. We can see that the average distances vary even less when this method is applied. In the 1700s, the sense pairs seen closest together are sight-sight, hear-sight, and hear-hear. It appears that the smell-taste and taste-taste pairings seen in Figure 4.4 are less strong. In the 1800s, the mixed modal sense pairs with the least distance are hear-sight, sight-smell, and hear-touch. This falls more in line with what was seen in Figure 4.4. Perhaps what's even more interesting is that the smell-taste grouping has the highest average distance in this period. The 1900s bar plot shows that sight-sight, hear-sight, and sight-touch are the closest in proximity. Finally, the full corpus sense pair ordering seems similar to the one for the 1800s except touch-touch and hear-taste are switched.

4.3.2 Total Sense Pairs In Radius Of 30

In addition to the average distance between sense descriptors, I was also interested in learning more about the pairs of senses that appeared within a radius of 30 in the generated sensory space. The radius was chosen based on previous work which determined this to be a good value from the average distances seen in this Project Gutenberg corpus [2]. Using

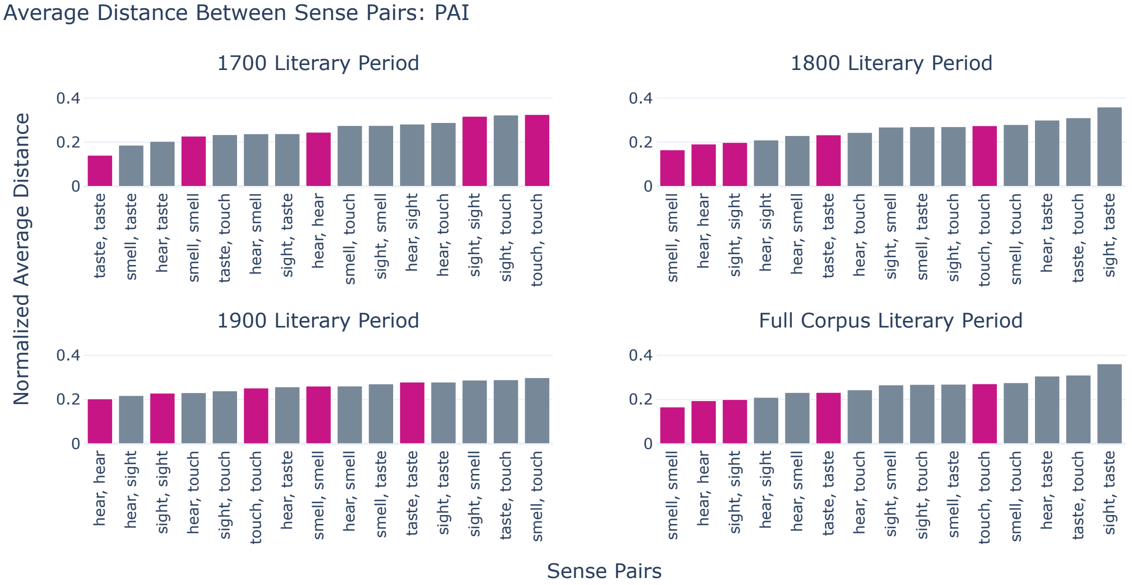

Figure 4.4: Each plot displays the normalized average distance between top 500 descriptors ranked using PAI. Descriptors belonging to the same sense are shown in magenta. Descriptors belonging to different senses are shown in gray.

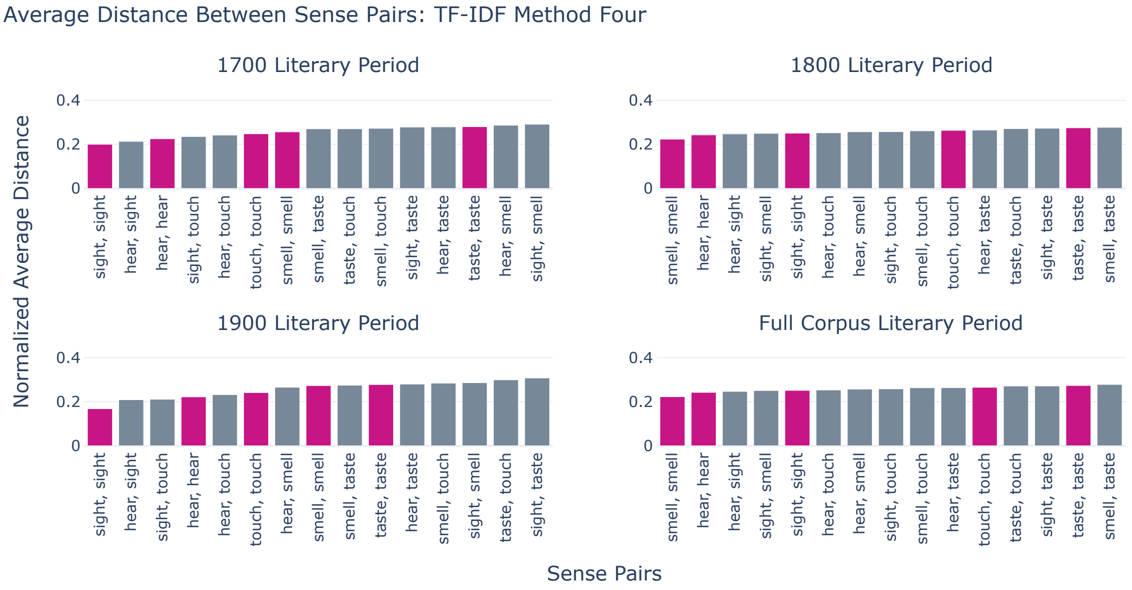

Figure 4.5: Each plot displays the normalized average distance between top 500 descriptors ranked using TF-IDF method four. Descriptors belonging to the same sense are shown in magenta. Descriptors belonging to different senses are shown in gray.

each individual descriptor in the top 500 descriptors, I looked to see what other descriptors occurred within the set radius of 30. As with the average distance between descriptors, I looked at the top 500 descriptors ranked using PAI and TF-IDF method four. The results for PAI are shown in Figure 4.7. The 1800s, 1900s, and the full corpus plots show that the hear-sight combination occurs most frequently in a radius of 30. This pattern has also been observed in the average distance plots for the 1800s and 1900s. In the 1700s, the pair with the greatest frequency is taste-touch followed by smell- taste. I also want to highlight that all five same modal pairs occur the least frequently within a radius of 30 in the 1700s and 1800s literary periods. Moving onto the TF-IDF method four results shown in Figure 4.7, the hear-sight pairing appears the most frequently in the 1800s and the full corpus, and the second most frequently in the 1900s. The sight-touch pairing is also in the top two for the 1800s and 1900s. In the 1700s, the smell-taste pair reemerges as the most frequent followed closely with smell-smell and taste-taste. This has been a reoccurring theme for the 1700s period as supported by my discussion on the average distance between sense pairs.

4.4 SENSORY BLENDS - OVERLAPPING DESCRIPTORS

In addition to the sensory pairs that are seen in these fiction texts, I wanted to see what the top descriptors were when applying PAI and TF-IDF method four. Examining the PCA graphs gave me a sense of the sensory landscape, but did not allow me to visualize the overlapping descriptors as easily. Thus, this section presents the top overlapping descriptors from the top 500 descriptors ranked with PAI and TF-IDF method four. The result of this exercise shows the frequency of the descriptor normalized by the size of the context window as seen in Figure 4.8 and Figure 4.9.

Looking at the overlapping descriptors identified using PAI, it is clear that there is no word that is common between all 3 literary periods. However, *tip* appears in both the 1700s and the 1800s descriptors sets. *Tip* is associated with smell and touch in both of these periods and is more frequently associated with smell than touch. This finding is rather unusual because *tip* does not naturally evoke depictions of the senses. Further examination of the the term's usage reveals that tip was used to refer to "the tip of the nose" in olfactory contexts and as "finger tip" in tactile contexts. Both uses refer to body parts associated with the senses. Overall, the 1700s literary period has the a smallest number of overlapping descriptors (only 4) compared to the other periods. One could attribute this small size to the fact that this corpus had the least number of texts. On the other hand, the 1800s and 1900s provide many more overlapping descriptors. The 1800 plot includes three version of the word smell: *smelled, smelling, and smell.* All three of these descriptors were associated

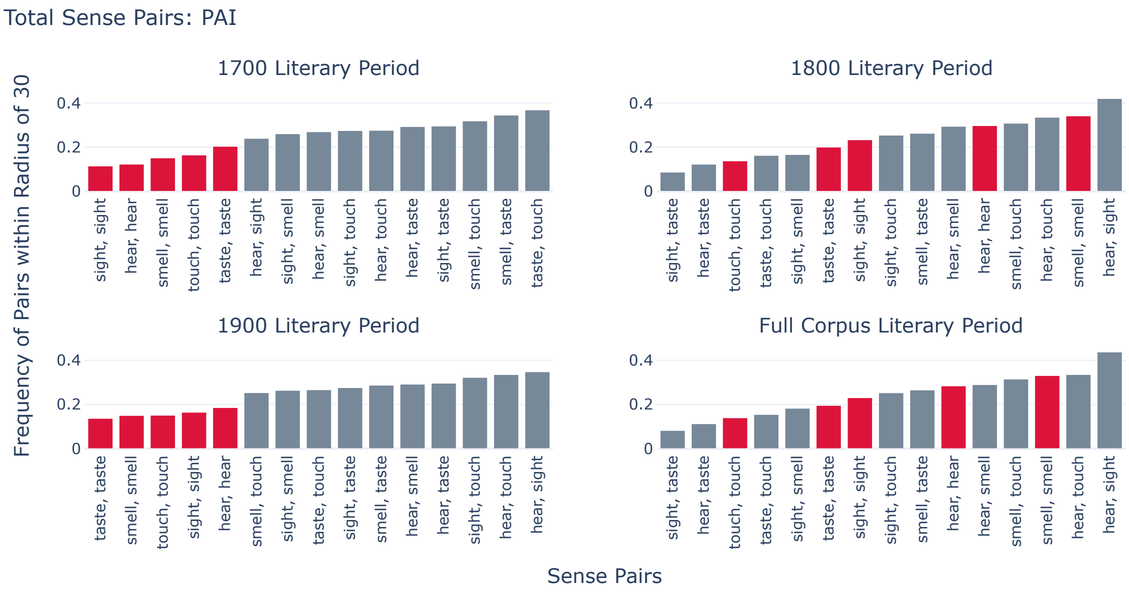

Figure 4.6: Sensory pairs within a radius of 30 around each descriptor in the set of top 500 descriptors ranked using PAI. Cross-modality pairs are shown in gray. Same-modality pairs are shown in red.

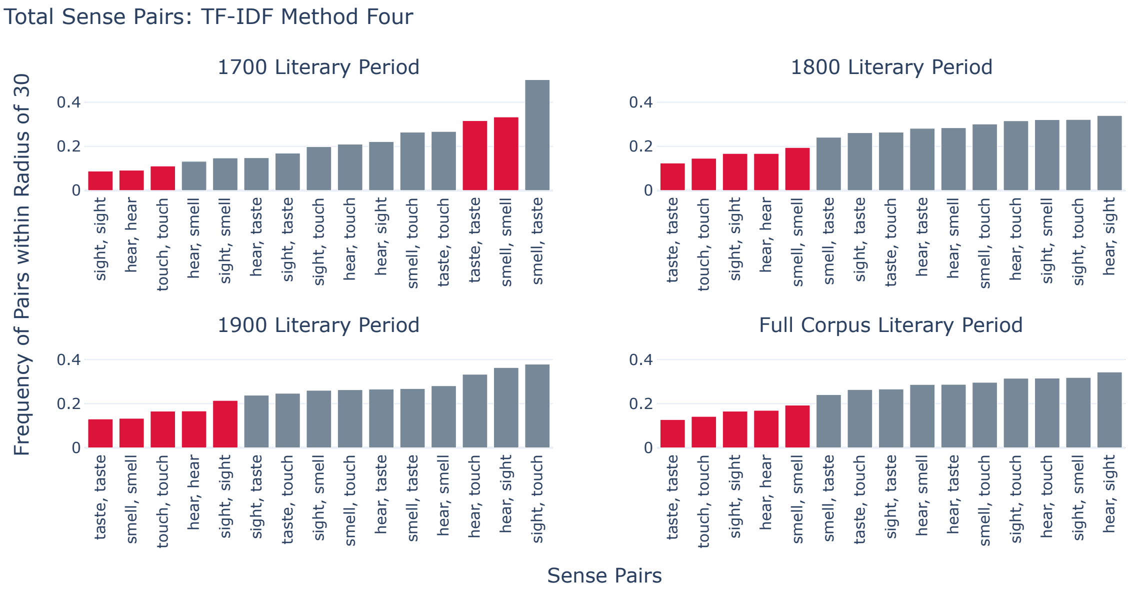

Figure 4.7: Sensory pairs within a radius of 30 around each descriptor in the set of top 500 descriptors ranked using TF-IDF method four. Cross-modality pairs are shown in gray. Same-modality pairs are shown in red.

with smell and taste. When looking at the 1900s literary period, it becomes apparent that taste descriptors make up a large part of the top overlapping senses. The majority of these taste descriptors are also associated with smell. In addition, words like *stifle* and *tingling* appear in smell and touch contexts.

I did a similar analysis of the overlapping descriptors using TF-IDF as seen in Figure 4.9. The resulting graphs show a very different picture from what was seen in the top descriptors for PAI. Here there is much more sensory blending as many of the modalities share the descriptors. I also noted that *said* was the most frequent descriptor in all 4 corpora and spanned 4-5 modalities. Other shared words included *little, man, time, hand, good, come, eyes, hand/hands, moment, let.* It's important to note that these words seem to be much more common than what was observed with the PAI descriptors. One reason the sixth document of non-sensory context windows was added to the TF-IDF method four was penalize those words that occurred frequently. It appears that more general words like *said*, for example, were not penalized enough. However, among the 500 descriptors, these were the only words identified to be overlapping so it is possible that the remaining words provide a robust representation of each sensory space. At the same time, it is important to take note that sensory descriptors can be associated with many different modalities (even all five).

4.5 DISCUSSION

In this chapter, I discuss the experimental results from PCA analysis of the sensory space and three additional metrics on the sensory descriptors extracted from this corpus. The 1700s period was the smallest data set looked at in this study. Examining the sensory spaces in this period was the most difficult as most descriptors were densely packed together and overlapping. This leads me to conclude that a larger data set will be needed to get the full picture of this sensory space. The only exception to this seen in the TF-IDF method four PCA plots which show a dispersed cloud of descriptors for the 1700s. The smell-taste interactions in this period plot were the most obvious. It was noted from all the PCA plots for the 1800s that some clusters of senses are more prevalent than others. **Sight-touch-hearing** formed one group and **smell-touch-hearing** formed another group. Parts of these groupings were also seen as sense pairs with the smallest average distance (**hear-sight and hear-touch and sight-touch**). Perhaps what is more intuitive is the **taste-smell** cluster, seen in some 1700s plots. Smell-taste sense pairs were also popular in the 1700s literary period. These same groupings were also seen in the PAI PCA plots for the 1900s literary period, further validating the idea that these are more general groupings rather than ones

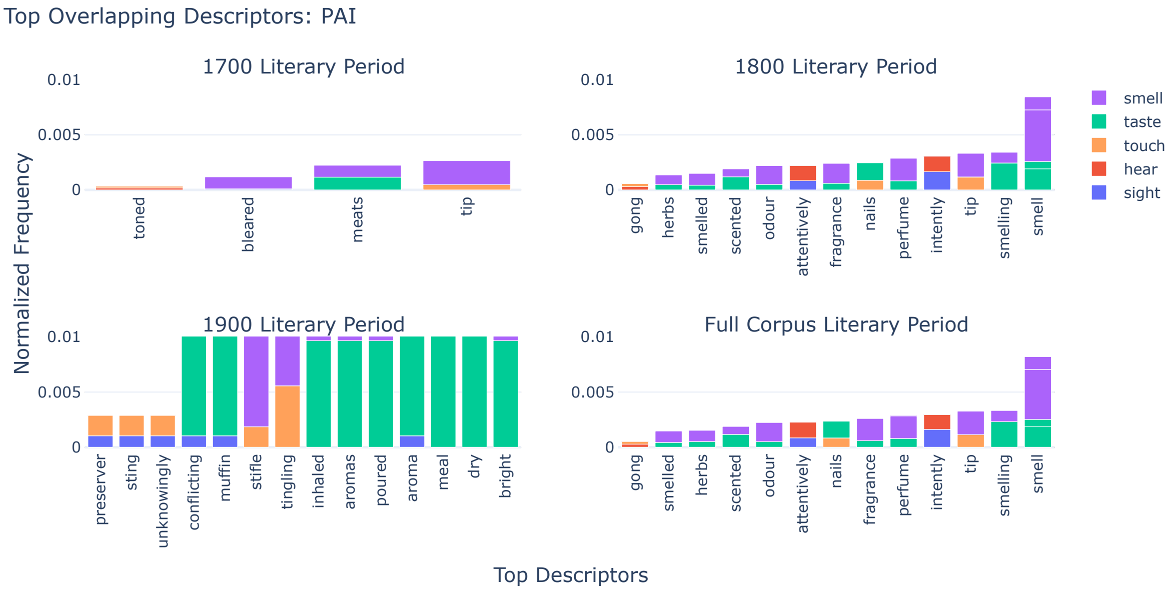

Figure 4.8: Top overlapping descriptors ranked with PAI

that can be attributed to a particular literary period.

It was evident from looking at the plots for all literary periods that there were no solitary senses. Cross-modal interactions within the same context were more common than not. Looking at the individual descriptors, it can be confirmed that descriptors tied to emotions are commonly associated with sight. Some plots also showed that taste descriptors were related to food and smell descriptors were olfactory descriptors of food. Besides these literal usages, examples of figurative descriptors were also seen. These descriptors were more scattered in the plots since the word embedding model did not identify them to be semantically similar.

This study also seeks to understand how effective each ranking approach is in identifying the top sensory descriptors. When comparing the different approaches used in this study, a few strengths and weaknesses stand out. PAI, for example, is an established technique for extracting sensory descriptors, but it causes infrequent descriptors to be ranked higher in small datasets. This problem was not observed in the larger data sets like the 1800s. Surprisingly, TF-IDF method two produced plots that were visually similar to the PAI plots with the exception of the 1900s literary period. The 1900s literary period was less defined and harder to cluster when this method was used. At the same time, this method proved to be very effective at eliminating descriptors that occurred infrequently in the dataset. Finally, TF-IDF method four demonstrated that a normalization step in the ranking process can change the orientation of the entire sensory space significantly. It should be noted that

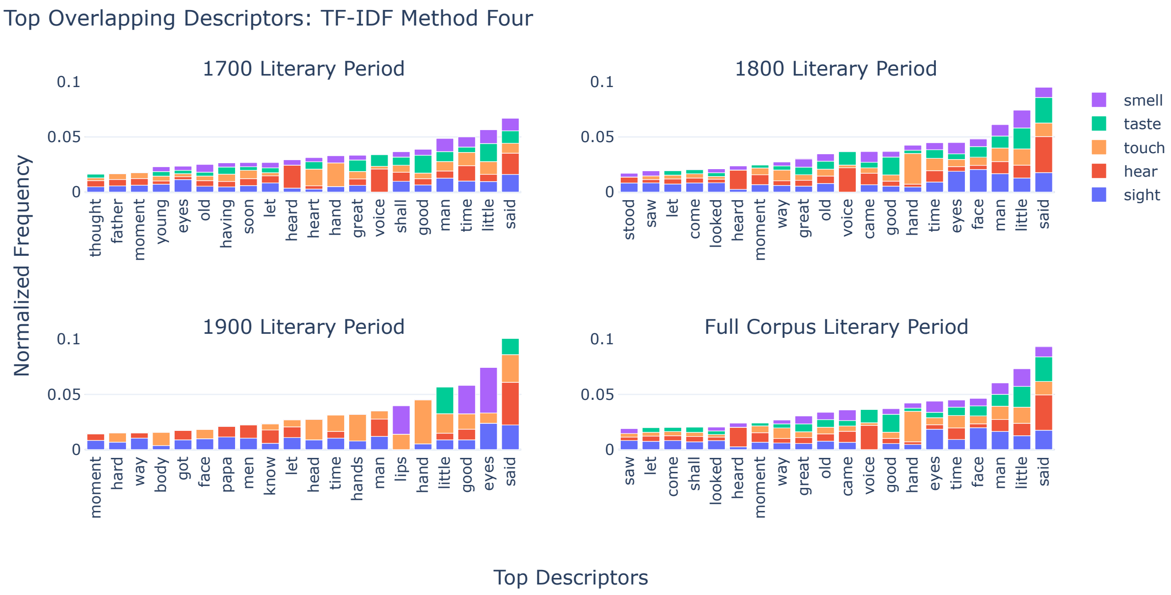

Figure 4.9: Top overlapping descriptors ranked with TF-IDF Method Four

the sensory space for the 1700s literary space changed the most when this technique was applied. It can be theorized that this technique yields more interpretable visualizations for a smaller data set. In addition, TD-IDF brought up more overlapping sensory descriptors than PAI. This observation could indicate that TF-IDF places a higher weight on frequent words regardless of whether they appear within the context windows of other sensory modalities. This study reports on the different organizations of sensory spaces created from these ranking methods, but it does not provide conclusive evidence that one ranking method is superior. It is suggested that a variety of ranking methods should be tested on a new dataset or application. Furthermore, the accuracy of the sensory spaces discovered in this study need to be verified with the help of linguists and other domain experts.

The approaches defined in this study allow us to visualize the sensory space in an interactive manner, but these approaches were not sufficient in identifying how literary movements may influence the use of sensory ideas. For example, overlapping descriptors from each literary period suggest that there is more in common between the periods than previously thought. These results could be evidence that sensory language does not evolve with literary movements. However, I hesitate to make this claim given the lack of evidence to support it. Due to time constraints, this study could not be reformed so that texts are grouped by country of origin. This may have helped to draw out the effects of literary movements. Further iterations of these approaches may be able to illustrate the sensory space more clearly and accurately in order to distill the specific literary movements.

CHAPTER 5: FUTURE WORK

So far, I have provided the details on my investigation into how the basic sensory modalities of sight, hearing, smell, taste, and touch interact in a corpus of fictional texts. One of the goals of this work was to apply existing tools in computation linguistics and natural language processing to better understand and interpret this data. In this chapter, I delve into possible future works.

5.1 RANKING METHODS

This study has used ranking to find the top descriptors from a set of context windows by applying four different ranking approaches: perception association index (PAI), TF-IDF (3 variations). The latter is commonly seen in search engine algorithms and document ranking problems. Other extensions of this work might benefit from exploring approaches outside the vector space model. One approach that has been previously studied in information retrieval is the Okapi-BM2 ranking function [49]. This widely used function is a probabilistic retrieval model that incorporates term frequency and document length. Related to the BM2 approach is LambdaRank, a machine learning approach to BM25 that uses the input attributes of BM25 [50, 51]. Previous studies have also simply ranked descriptors using overall word frequency in the corpus [2]. Following this line of thinking, it would be interesting to observe the top descriptors using the frequency of the word in context window set.

5.2 COMPUTATIONAL MODELS

The novelty of this work lies in how it uses computation models to map out the sensory space. I used a word2vec model with CBOW model on the contexts windows using a hidden layer with 200 units, minimum word count as 1, and 30 training iterations. Of course, the parameters of this model can be changed and should be tested for different data sets and domains. Furthermore, each model was dependent of the number of texts available for a literary period. For example, the 1800s literary period produced much more context windows that the the 1700s or 1900s periods. It's important to investigate the sensitivity of these models to these parameters. In this study, I have used to gensim library to create my word embeddings, but others like spacy should be tested to determine which will provide the best results for this data set [39, 52, 53]. While word embeddings are a good start, these models don't capture the full context. Sensory language relies on more than individuals words and can take on figurative interpretations. In fact, metaphors provide a mechanism for humans to understand the physical and social world we live in [54]. Metaphor and similes appear more commonly in the form of phrases and can be explored using phrase embeddings or more complex models like BERT [55].

5.3 UNSUPERVISED LEARNING

PCA plots used to visually observe how sensory descriptors interact and connect. While they provides key clues as to how sensory spaces are organized, understanding the clusters of descriptors and senses was a tedious and manual process. To identify the clusters of data in a empirical manner, future work could used unsupervised learning approaches to identify the sensory groupings. For example, K-mean clustering has been used previously for topic classification in environmental education journals [56]. Furthermore, k-means clustering has been used in conjunction with PCA to analyze text data on social media [57]. Another clustering technique is Agglomerative Nesting (AGNES) hierarchical clustering [58]. AGNES has been previously used to explore sensory spaces of English perceptual verbs [14].

5.4 TOPIC MODELLING

In addition to supervised learning, sensory spaces can be re-imagined as topics in a corpus. Topic modelling is heavily studied in natural language processing [51, 56]. Rather than studying each text, a topic modelling approach might use a set of sensory contexts to extract the higher level topics. From the standpoint of understanding linguistic behavior in different literary periods, topic modelling might help us to understand the overall themes in a period in an empirical manner. This work could assist linguistics who are trying to map out the sociopolitical topics discussed in each literary period. One possible approach for topic modelling is latent Dirichlet allocation. This generative statistical model aims to find the topics that a document belongs to based on the words contained in the document [51]. Many variations of LDA have appeared over the years and have been evaluated for different applications [59]

5.5 DATA SETS

For this study, I utilized the Gutenberg dammit published by Allison Parrish. One primary reason for using this data is that it is open-source and it contains meta-data which allows me to connect each text to a literary period. However, I want to highlight that this data set did not provide me with an equal distribution of texts across all literary periods as shown in Table 3.1. Finding a large enough data set can be a barrier for natural language processing tasks. Future works should consider augmenting this data set with other fiction texts from Project Gutenberg. Alternatively, other open-source fiction data sets can be studied [60, 61]. Fiction texts were studied due to the copious amount of sensory expressions present. Other types of text data sources, like electronic medical records, telehealth transcripts, social media posts, film subtitles, poetry, and transcripts from chat bot conversations, can provide fascinating insights into how sensory language is used in modern day applications.

5.6 SEED WORDS

Finally, this study's context windows and descriptors rely heavily on the quality of the seed words used to extract them. To understand the impact of these seed words, additional analysis should be conducted. While I used a semi-automatic approach to create the set of seed words, this approach should be polished using linguistic domain knowledge. Other dictionaries and ontologies should be tested. In fact, previous works have studied the sensorimotor strength norms of English words [62]. Existing works on the sensory, perceptual landscape of English words can provide a basis for building this list of seed words.

CHAPTER 6: CONCLUSION

Sensory experiences shape the world that we live in. They are ingrained in how we perceive the world and create the landscape of our mind. Sensory language, in particular, gives us the ability to empathize with characters and situations. This study reports on the connections between different sensory modalities and emotions presented in English fiction using a data-driven model. I extracted sensory context windows from a data set of Project Gutenberg books separated into three literary periods: 1700s, 1800s, and 1900s. This approach uses distributional-semantic word embeddings to map semantically related sensory descriptors. My study is also unique in how it experiments with ranking algorithms to identify the top descriptors in each period. The results of this study are presented in the form of interactive, visual graphs. The findings show that certain sensory grouping may be more common than others. Further iterations of this work will be useful in developing adaptive and inclusive technologies, understanding sensory systems, and strengthening creative writing.

REFERENCES

[1] K. Bonna, K. Finc, M. Zimmermann, Bola, P. Mostowski, M. Szul, P. Rutkowski, W. Duch, A. Marchewka, K. Jednoróg, and M. Szwed, “Early deafness leads to re-shaping of global functional connectivity beyond the auditory cortex,” 2019. [Online]. Available: https://arxiv.org/abs/1903.11915

[2] R. Girju and C. Lambert, “Inter-sense: An investigation of sensory blending in fiction,” *CoRR*, vol. abs/2110.09710, 2021. [Online]. Available: https://arxiv.org/abs/2110.09710

[3] A. Ashokan, “Exploring the senses and emotions and their interconnections throughout literary periods,” Apr 2022. [Online]. Available: osf.io/r9aw4

[4] L. Zunshine, “Why we read fiction: Theory of mind and the novel,” 2006.

[5] B. C. Southgate, “History meets fiction,” 2009.

[6] B. Winter, *Sensory Linguistics: Language, perception and metaphor*, ser. Converging Evidence in Language and Communication Research. John Benjamins Publishing Company, 2019. [Online]. Available: https://books.google.com/books?id=FnGQDwAAQBAJ

[7] *The Varieties of sensory experience : a sourcebook in the anthropology of the senses / edited by David Howes.*, ser. Anthropological horizons [1]. Toronto ;: University of Toronto Press, 1991.

[8] P. L. Duffy, “Synesthesia in literature,” 2013.

[9] O. Deroy, *Sensory Blending: On Synaesthesia and related phenomena.* OUP Oxford, 2017. [Online]. Available: https://books.google.com/books?id=H161DgAAQBAJ

[10] G. O’Malley, “Literary synesthesia,” *The Journal of Aesthetics and Art Criticism*, vol. 15, no. 4, pp. 391–411, 1957. [Online]. Available: http://www.jstor.org/stable/427153

[11] N. Ruddick, “”synaesthesia” in emily dickinson’s poetry,” *Poetics Today*, vol. 5, no. 1, pp. 59–78, 1984. [Online]. Available: http://www.jstor.org/stable/1772426

[12] J. Huysmans, M. Mauldon, N. White, L. White, and F. Mauldon, *Against Nature*, ser. Oxford world’s classics. Oxford University Press, 1998. [Online]. Available: https://books.google.com/books?id=bPQgRK7D3i0C

[13] T. Parker, *The Fallen LP.* HarperCollins, 2007. [Online]. Available: https://books.google.com/books?id=yD4owOKThSIC

[14] R. Girju and D. Peng, "Exploring the sensory spaces of english perceptual verbs in natural language data," *CoRR*, vol. abs/2110.09721, 2021. [Online]. Available: https://arxiv.org/abs/2110.09721

[15] R. Brate, P. Groth, and M. van Erp, "Towards olfactory information extraction from text: A case study on detecting smell experiences in novels," *CoRR*, vol. abs/2011.08903, 2020. [Online]. Available: https://arxiv.org/abs/2011.08903

[16] G. Iatropoulos, P. Herman, A. Lansner, J. Karlgren, M. Larsson, and J. K. Olofsson, "The language of smell: Connecting linguistic and psychophysical properties of odor descriptors," *Cognition*, vol. 178, pp. 37–49, 2018. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0010027718301276

[17] T. Hörberg, M. Larsson, and J. Olofsson, "Mapping the semantic organization of the english odor vocabulary using natural language data," Apr 2020. [Online]. Available: psyarxiv.com/hm8av

[18] S. Tonelli and S. Menini, "FrameNet-like annotation of olfactory information in texts," in *Proceedings of the 5th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*. Punta Cana, Dominican Republic (online): Association for Computational Linguistics, Nov. 2021. [Online]. Available: https://aclanthology.org/2021.latechclfl-1.2 pp. 11–20.

[19] M. Klarer, *An introduction to literary studies*. Routledge, 2013.

[20] Z. Xiaofen, "On the influence of naturalism on american literature," *English Language Teaching*, vol. 3, 05 2010.

[21] D. R. Amancio, O. N. Oliveira, and L. da Fontoura Costa, "Identification of literary movements using complex networks to represent texts," *New Journal of Physics*, vol. 14, no. 4, p. 043029, apr 2012. [Online]. Available: https://doi.org/10.1088/1367-2630/14/4/043029

[22] B. Wallace, "Multiple narrative disentanglement: Unraveling infinite jest," in *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Montréal, Canada: Association for Computational Linguistics, June 2012. [Online]. Available: https://aclanthology.org/N12-1001 pp. 1–10.

[23] D. Elson, N. Dames, and K. McKeown, "Extracting social networks from literary fiction," in *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*. Uppsala, Sweden: Association for Computational Linguistics, July 2010. [Online]. Available: https://aclanthology.org/P10-1015 pp. 138–147.

[24] P. Jayannavar, A. Agarwal, M. Ju, and O. Rambow, "Validating literary theories using automatic social network extraction," in *Proceedings of the Fourth Workshop on Computational Linguistics for Literature*. Denver, Colorado, USA: Association for Computational Linguistics, June 2015. [Online]. Available: https://aclanthology.org/W15-0704 pp. 32–41.

[25] H. He, D. Barbosa, and G. Kondrak, “Identification of speakers in novels,” in *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Sofia, Bulgaria: Association for Computational Linguistics, Aug. 2013. [Online]. Available: https://aclanthology.org/P13-1129 pp. 1312–1320.

[26] A. van Cranenburgh and C. Koolen, “Identifying literary texts with bigrams,” in *Proceedings of the Fourth Workshop on Computational Linguistics for Literature*. Denver, Colorado, USA: Association for Computational Linguistics, June 2015. [Online]. Available: https://aclanthology.org/W15-0707 pp. 58–67.

[27] T. Schmidt, K. Dennerlein, and C. Wolff, “Emotion classification in German plays with transformer-based language models pretrained on historical and contemporary language,” in *Proceedings of the 5th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*. Punta Cana, Dominican Republic (online): Association for Computational Linguistics, Nov. 2021. [Online]. Available: https://aclanthology.org/2021.latechclfl-1.8 pp. 67–79.

[28] K. Clark, M.-T. Luong, Q. V. Le, and C. D. Manning, “Electra: Pre-training text encoders as discriminators rather than generators,” 2020. [Online]. Available: https://arxiv.org/abs/2003.10555

[29] J. Brooke, A. Hammond, and G. Hirst, “GutenTag: an NLP-driven tool for digital humanities research in the Project Gutenberg corpus,” in *Proceedings of the Fourth Workshop on Computational Linguistics for Literature*. Denver, Colorado, USA: Association for Computational Linguistics, June 2015. [Online]. Available: https://aclanthology.org/W15-0705 pp. 42–47.

[30] A. Hammond, J. Brooke, and G. Hirst, “A tale of two cultures: Bringing literary analysis and computational linguistics together,” in *Proceedings of the Workshop on Computational Linguistics for Literature*. Atlanta, Georgia: Association for Computational Linguistics, June 2013. [Online]. Available: https://aclanthology.org/W13-1401 pp. 1–8.

[31] I. Rakkolainen, A. Farooq, J. Kangas, J. Hakulinen, J. Rantala, M. Turunen, and R. Raisamo, “Technologies for multimodal interaction in extended realitymdash;a scoping review,” *Multimodal Technologies and Interaction*, vol. 5, no. 12, 2021. [Online]. Available: https://www.mdpi.com/2414-4088/5/12/81

[32] U. B. Qushem, A. Christopoulos, S. S. Oyelere, H. Ogata, and M.-J. Laakso, “Multimodal technologies in precision education: Providing new opportunities or adding more challenges?” *Education Sciences*, vol. 11, no. 7, 2021. [Online]. Available: https://www.mdpi.com/2227-7102/11/7/338

[33] E. Schreuder, J. van Erp, A. Toet, and V. Kallen, “Emotional responses to multisensory environmental stimuli: A conceptual framework and literature review.” *SAGE Open*, vol. 6, no. 1, pp. –, Feb. 2016.

[34] A. Mills, “B. landau amp; l. r. gleitman, language and experience. evidence from the blind child. cambridge ma: Harvard u.p., 1985. pp. xi 250.” *Journal of Child Language*, vol. 14, no. 2, p. 397–402, 1987.

[35] R. Plutchik, “The nature of emotions: Human emotions have deep evolutionary roots, a fact that may explain their complexity and provide tools for clinical practice,” *American Scientist*, vol. 89, no. 4, pp. 344–350, 2001. [Online]. Available: http://www.jstor.org/stable/27857503

[36] E. Clark, A. S. Ross, C. Tan, Y. Ji, and N. A. Smith, “Creative writing with a machine in the loop: Case studies on slogans and stories,” in *23rd International Conference on Intelligent User Interfaces*, ser. IUI ’18. New York, NY, USA: Association for Computing Machinery, 2018. [Online]. Available: https://doi.org/10.1145/3172944.3172983 p. 329–340.

[37] O. Bullock, “Poetry and trauma: exercises for creating metaphors and using sensory detail,” *New Writing*, vol. 18, no. 4, pp. 409–420, 2021. [Online]. Available: https://doi.org/10.1080/14790726.2021.1876094

[38] L. Kelly, “Playing with the diary: how crafting a multimodal and sensory diary can have a positive impact on teacher wellbeing,” *Reflective Practice*, vol. 23, no. 1, pp. 1–16, 2022. [Online]. Available: https://doi.org/10.1080/14623943.2021.1973986

[39] M. Honnibal, I. Montani, S. Van Landeghem, and A. Boyd, “spaCy: Industrial-strength Natural Language Processing in Python,” 2020.

[40] S. Bird, E. Klein, and E. Loper, *Natural Language Processing with Python: Analyzing Text with the Natural Language Toolkit.* O’Reilly Media, Inc., 2009.

[41] S. Jabri, A. Dahbi, T. Gadi, and A. Bassir, “Ranking of text documents using tf-idf weighting and association rules mining,” in *2018 4th International Conference on Optimization and Applications (ICOA)*, 2018, pp. 1–6.

[42] J. Shlens, “A tutorial on principal component analysis,” *CoRR*, vol. abs/1404.1100, 2014. [Online]. Available: http://arxiv.org/abs/1404.1100

[43] M. R. Mahmoudi, M. H. Heydari, S. N. Qasem, A. Mosavi, and S. S. Band, “Principal component analysis to study the relations between the spread rates of covid-19 in high risks countries,” *Alexandria Engineering Journal*, vol. 60, no. 1, pp. 457–464, 2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1110016820304543

[44] S. Sehgal, H. Singh, M. Agarwal, V. Bhasker, and Shantanu, “Data analysis using principal component analysis,” in *2014 International Conference on Medical Imaging, m-Health and Emerging Communication Systems (MedCom)*, 2014, pp. 45–48.

[45] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[46] V. L. Kinner, L. Kuchinke, A. M. Dierolf, C. J. Merz, T. Otto, and O. T. Wolf, "What our eyes tell us about feelings: Tracking pupillary responses during emotion regulation processes," *Psychophysiology*, vol. 54, no. 4, pp. 508–518, 2017. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1111/psyp.12816

[47] S. Shimojo and L. Shams, "Sensory modalities are not separate modalities: plasticity and interactions," *Current Opinion in Neurobiology*, vol. 11, no. 4, pp. 505–509, 2001. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0959438800002415

[48] E. Wehling, "The neuropsychology of smell and taste," *Neuropsychological Rehabilitation*, vol. 24, no. 5, pp. 807–808, 2014. [Online]. Available: https://doi.org/10.1080/09602011.2014.908540

[49] S. E. Robertson and S. Walker, "Some simple effective approximations to the 2-poisson model for probabilistic weighted retrieval," in *Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, ser. SIGIR '94. Berlin, Heidelberg: Springer-Verlag, 1994, p. 232–241.

[50] K. M. Svore and C. J. Burges, "A machine learning approach for improved bm25 retrieval," in *Proceedings of the 18th ACM Conference on Information and Knowledge Management*, ser. CIKM '09. New York, NY, USA: Association for Computing Machinery, 2009. [Online]. Available: https://doi.org/10.1145/1645953.1646237 p. 1811–1814.

[51] C. Zhai and S. Massung, *Text Data Management and Analysis: A Practical Introduction to Information Retrieval and Text Mining.* Association for Computing Machinery and Morgan amp; Claypool, 2016.

[52] R. Rehurek and P. Sojka, "Gensim–python framework for vector space modelling," *NLP Centre, Faculty of Informatics, Masaryk University, Brno, Czech Republic*, vol. 3, no. 2, 2011.

[53] P. Bojanowski, E. Grave, A. Joulin, and T. Mikolov, "Enriching word vectors with subword information," *arXiv preprint arXiv:1607.04606*, 2016.

[54] J. M. Lawler, *Language*, vol. 59, no. 1, pp. 201–207, 1983. [Online]. Available: http://www.jstor.org/stable/414069

[55] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding." [Online]. Available: https://arxiv.org/abs/1810.04805

[56] I.-C. Chang, T.-K. Yu, Y.-J. Chang, and T.-Y. Yu, "Applying text mining, clustering analysis, and latent dirichlet allocation techniques for topic classification of environmental education journals," *Sustainability*, vol. 13, no. 19, 2021. [Online]. Available: https://www.mdpi.com/2071-1050/13/19/10856

[57] M. M. J. Adnan, M. L. Hemmje, and M. A. Kaufmann, "Social media mining to study social user group by visualizing tweet clusters using word2vec, PCA and k-means," in *Joint Proceedings of the Second Workshop on Bridging the Gap between Information Science, Information Retrieval and Data Science, and Third Workshop on Evaluation of Personalisation in Information Retrieval co-located with 6th ACM SIGIR Conference on Human Information Interaction and Retrieval (CHIIR 2021), Canberra, Australia (Virtual Event), March 19th, 2021*, ser. CEUR Workshop Proceedings, I. Frommholz, H. Liu, M. Melucci, N. J. Belkin, G. J. F. Jones, N. Kando, and G. Pasi, Eds., vol. 2863. CEUR-WS.org, 2021. [Online]. Available: http://ceur-ws.org/Vol-2863/paper-05.pdf pp. 40–51.

[58] L. Kaufman and P. J. Rousseeuw, *Finding Groups in Data: An Introduction to Cluster Analysis.* John Wiley, 1990.

[59] U. Chauhan and A. Shah, "Topic modeling using latent dirichlet allocation: A survey," *ACM Comput. Surv.*, vol. 54, no. 7, sep 2021. [Online]. Available: https://doi.org/10.1145/3462478

[60] A. Piper, "Fictionality," 2016. [Online]. Available: https://doi.org/10.7910/DVN/5WKTZV

[61] A. Piper, "txtLAB Contemporary Novel Data Set," 1 2016. [Online]. Available: https://figshare.com/articles/dataset/_txtLAB_Contemporary_Novels/2061990

[62] D. Lynott, L. Connell, M. Brysbaert, J. Brand, and J. Carney, "The lancaster sensorimotor norms: multidimensional measures of perceptual and action strength for 40,000 english words," *Behavior Research Methods*, vol. 52, pp. 1271 – 1291, 2019.

APPENDIX A: PCA PLOTS

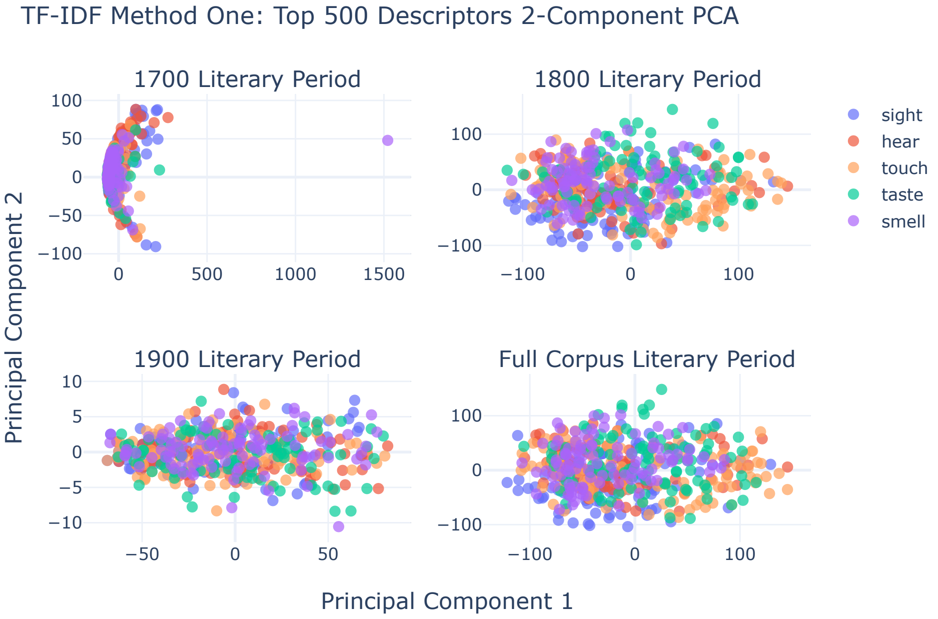

Figure A.1: PCA plot displaying the top 500 descriptors ranked using TF-IDF method one.

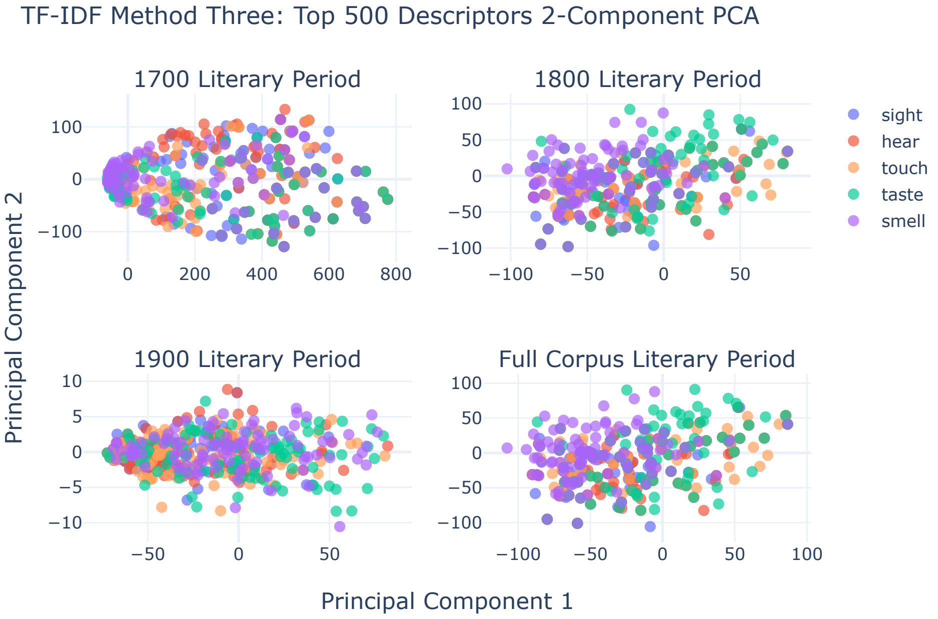

Figure A.2: PCA plot displaying the top 500 descriptors ranked using TF-IDF method three.