# Classifying Synesthetic Descriptions in Classic Literature

Data Mining Project

Ashling Scott, ascot105@kent.edu

March 31st, 2024

## Summary

**Synesthsia is the blending together of multiple senses. In literature, synesthesia can be used to create immersive, evocative descriptions, by engaging in readers multiple senses at once. This project aims to explore that technique, by analyzing classic English literature in search of instances of synesthesia. Using data mining and pattern matching, I intend to quantify the total counts of sensory description, and find out the rates at which both single sense and multi-sense descriptions occur.**

## 1. Introduction

Synesthesia is a perceptual phenomenon where sensory information of one type is perceived as a different type of sensory input. For instance, someone with synesthesia might associate certain colors with a scent, or may feel a tingling sensation when presented with certain sounds. Roughly four percent of the world population experiences synesthesia in one form or another.[2]

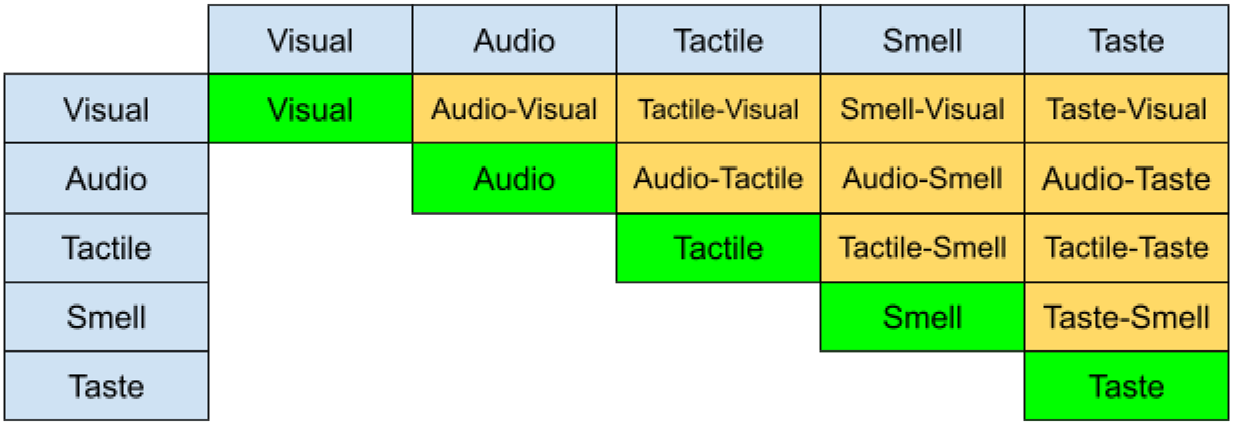
In literature, synesthesia can be leveraged to create compelling and novel descriptions. Expressions like “taste of victory” or “loud colors” combine senses in artistic ways, which add depth and intrigue to the prose. This creates an immersive and evocative experience for the reader, by engaging a variety of senses in a creative manner. Authors can create vivid and memorable experiences in this way.[3]

In this project, I will be analyzing classic literature to find examples of synesthetic descriptions. I hope to classify instances of synesthesia in class works, and identify which types of synesthesia are more common. Hopefully this will illuminate the history of using this literary technique, and show which associations between senses have been most deeply rooted in human history and arts.

## 2. Project Description

I will be analyzing the text of classic literature, searching for examples of synesthetic description, using this dataset:
https://www.kaggle.com/datasets/raynardj/classic-english-literature-corpus/data[1]
It contains over the full text of over 1000 classic literary works, dating back centuries. The text will be analyzed line by line, searching for sentence fragments that contain multiple sensory descriptions. These sensory descriptions will be within keyword for each sensory category, to be able to assign every pair of senses its own category.

Sentence fragments are used over full sentences (delineated by periods, commas, exclamation marks, and newline characters) because older literature tends to have very long sentences that have a high amount of commas. To avoid false positives where very long, rambling sentences happen to mention multiple senses, I've taken a more conservative approach and decided to look at sentence fragments instead. There are 15 categories in which I will place my sentence fragments, 5 single-sense categories, and 10 multi-sense. Sentence fragments may be placed in more than one category. I will analyse the rates that each sense shows up, as well as the rates that each multi-sense combination shows up, and look for patterns, such as where my expectations and the true counts differ.



| | Visual | Audio | Tactile | Smell | Taste |
|---|---|---|---|---|---|
| Visual | Visual | Audio-Visual | Tactile-Visual | Smell-Visual | Taste-Visual |
| Audio | | Audio | Audio-Tactile | Audio-Smell | Audio-Taste |
| Tactile | | | Tactile | Tactile-Smell | Tactile-Taste |
| Smell | | | | Smell | Taste-Smell |
| Taste | | | | | Taste |

The above table shows possible categories for pairs of sensory descriptors. The green categories are normal sensory descriptions, where sensory type corresponds to a single sense, and the yellow categories show types of synesthetic descriptions, pairs of 2 senses being used simultaneously. In my analysis, I expect to find many examples of green categories. What will be interesting is what yellow categories show up more often than others.

## 3. Background

For this project, I'm using using python for the scripting, and R for data visualization. For interacting with the database of texts, I'll be using the sqlite3 package in python. This allows us to use SQL commands to get information from or write to my SQLite database.

I was originally using the spaCy package, which is an NLP library for python that provides an easy way to tokenize text, and sophisticated pattern matching, but I found that this was unnecessary for r purpose in the end. The pattern matching, while more sophisticated, also adds unnmyeded complexity and a much longer runtime, so in the end I went with using regular expressions and splitting strings for tokenization, as well as a set of lists that represent keywords. This will allow us to easily identify which sentences contain descriptors from multiple sensory inputs.

I programmed the data mining script in python, using a functional programming approach. I processed over ten thousand text files, comprising over 1000 individual works of literature, and found sentences with sensory descriptors, then wrote my flagged totals into the database, allowing for easy analysis in the future.

I took the data into R to perform data visualization, using the ggplot2 library. I was most interested to look at the rates that senses appeared, both in the single-sense category as well as multi-sense fragments. I also used visualization to see the difference between expected multi-sense totals and actual, to see which pairs were combined more or less often than expected.

## 4. Problem Definition

Given a set of classic documents **D** $\{D_1, D_2, ...D_x\}$
And 5 sets of sensory descriptor words: **Vis**, **Aud**, **Tac**, **Sme**, and **Tas**
$\{Vis_1, Vis_2, ...Vis_x\}$, $\{Aud_1, Aud_2, ...Aud_x\}$, $\{Tac_1, Tac_2, ...Tac_x\}$, $\{Sme_1, Sme_2, ...Sme_x\}$, $\{Tas_1, Tas_2, ...Tas_x\}$

1. Tokenize each document from **D** down into sentences,
2. Search through those sentences to find instances where sensory descriptor words are present.
3. Classify each identified sentence into one or more of the 15 categories, given by the above table, based on what categories the matching words fall into.
4. Analyze distribution of classes to see what patterns can be found in the data. Identify which categories are most common, and what literature uses them the most often.

The biggest challenges of this project will be ensuring that I catch every example of synesthesia, and additionally in avoiding false positives. The pattern lists for each sense will need to be exhaustive, in order to account for the wide variety of sensor words in the english language. There may be some words that account for sensory perception that slip through the cracks, especially when analyzing older texts that use archaic words and sentence structures. If it proves too difficult to lock in on examples in classic literature, I may need to use a more modern set of literature to get enough data.

False positives could be an issue as well, by setting off the pattern recognition despite the sensory terms clearly being from two separate thoughts. Hopefully the measure of the distance between the two words can help alleviate this concern. Another way to address this issue may be to change the method of tokenization, and tokenize from both periods and commas, in order to break separate thoughts down into separate tokens.

After initially trying this project with full sentences, I eventually added commas to my splitting, therefore collecting sentence fragments.  This is because older literature tends to have very long sentences with many commas, which made it too easy for non-synesthetic matches to creep in.  I limited it to sentence fragments to better ensure that the two words from the lists are both used in quick succession, and thus more likely to be an instance of synesthesia.

### 5. Proposed Techniques

The database I selected had a slightly cumbersome way of storing the data I needed.  The text_file table had the full text off all the works, split up into over 10,000 blocks of text and ids, but no reference to the name of the work.  In order to get the full text of any particular piece of literature, I had to utilize a SQL table join.  I joined text_files and book_file tables together, to be able to get the id, name, and full text of the current book and make sure I had all the data in the right place.

```
query = ”””
        SELECT text
        FROM text_files natural join book_file
        WHERE book_file.book_id = ””” + str(x)
```

I used a regular expression to delineate my sentence fragments, splitting sentences on periods, commas, exclamation marks, and newline characters:

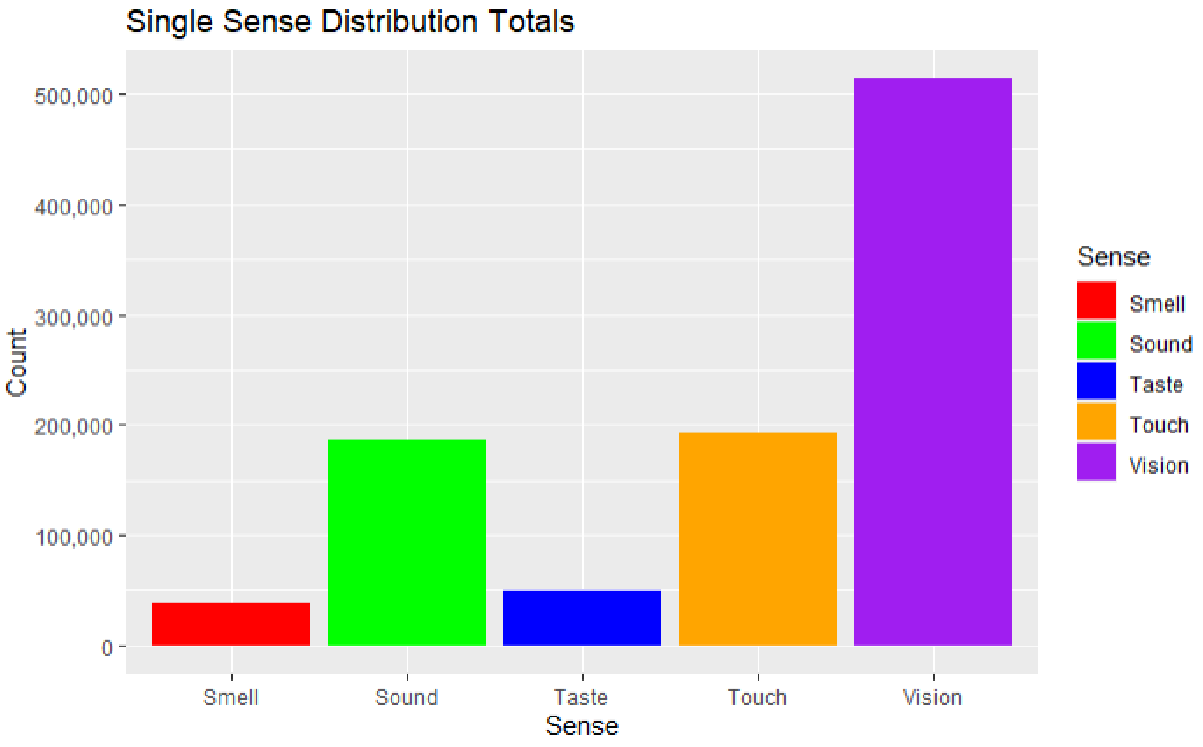
```
sentences = re.split(r"[.!\n,]+", text_content)
```

I analyzed every sentence fragment for key words, and set five boolean values (vision, sound, smell, taste, touch) to flag out sensory words that appeared in the fragment.  I then sorted the sentence into one or more lists of sentence fragments corresponding with each synesthetic combination.  Then I calculated the total number of fragments that had been sorted into each list and placed them into my totals table along with the book name and id, and then I wrote the totals list into a new SQL table created specifically for tracking synesthesetic instances.

In R, I imported data from the new table and did a quick bit of data wrangling to get the data into an acceptable dataframe, and then utilized ggplot2 to perform data visualization.  I focused on bar charts since the main interest of this project was in finding counts of occurrences, and I also looked at some lists of the most frequent and least frequent synesthetic counts among the literature analyzed.

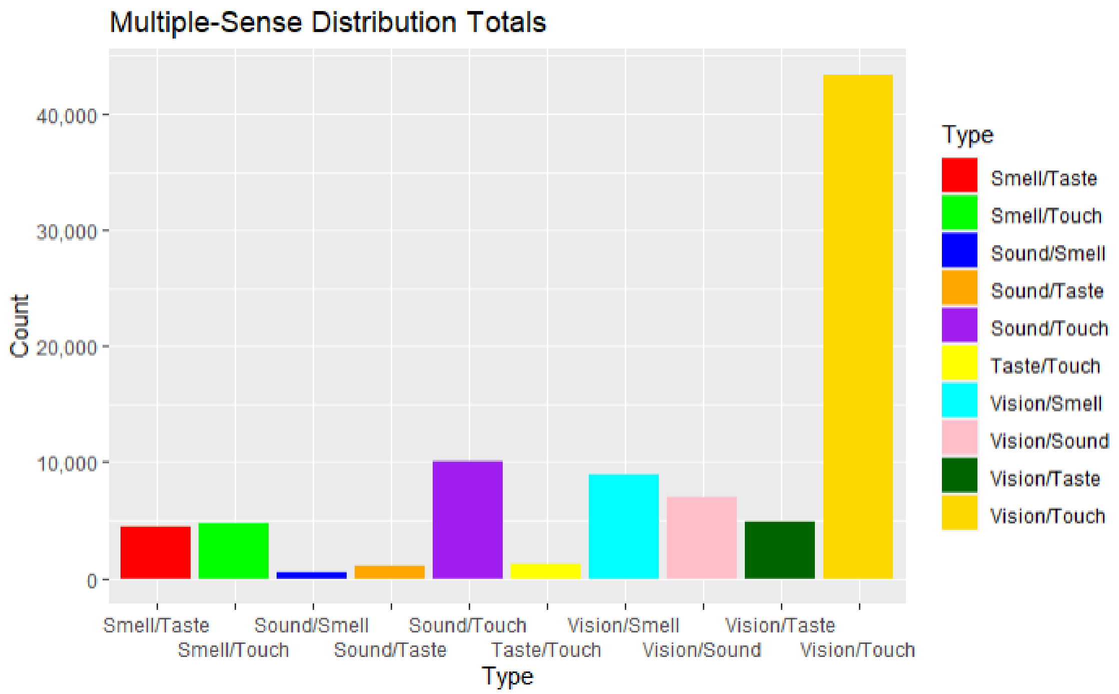
## 6. Visualization

The first graph is a count of single sensory instances. This offers a baseline for how often particular senses show up in classic English literature, whether they're alone or with other senses.

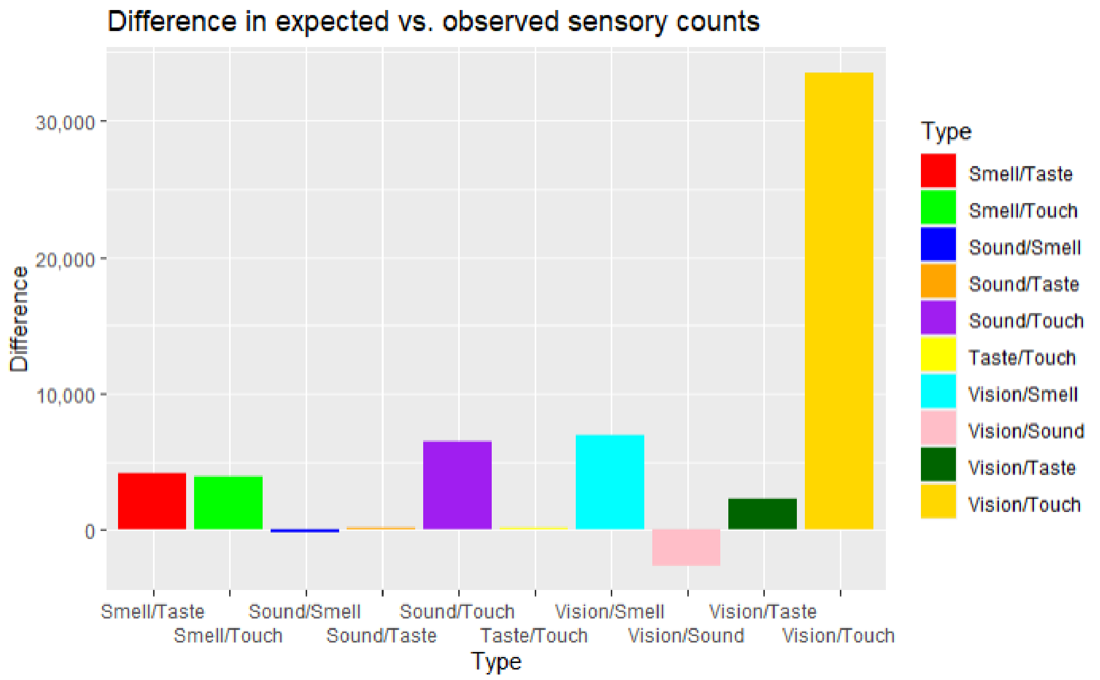


As one might expect, vision is the most prevalent sense, due to humans being a visual species. When descriptions occur in literature, visual descriptions are viewed as the most important. The next level of frequency is sound and touch, which both come in with levels just under 200,000. I didn't expect touch to be higher than sound, since audio is also an important sense for humans, and I didn't expect touch to translate as well into text. The least commonly used descriptive senses are smell and taste, which is expected.

Next we can take a look at multi-sense distributions. There are 10 categories in this graph, representing the 10 different combinations of 2 senses. If distribution of senses was independent, I would expect to see distributions with similar proportions to the single sense graph, with vision/sound and vision/touch being the highest. However, this isn't quite what I get.



As you can see, Vision/Touch dominates the multi-sense distribution graph, comprising nearly as much as all the other combinations combined. I would expect this to be the highest combination, but not to this extent. Meanwhile, Vision/Sound which was my initial prediction for the highest multi-sense is getting beaten by some unusual competitors: Sound/Touch and Vision/Smell. Something about the link between sound and vision words doesn't lend itself to forming sentence fragments.



I created this graph of differences between expected and observed multi-sense counts, in order to illustrate the difference. The expect values were calculated based on the rate of single-sense words appearing, and then multiplied under the assumption that the variables were independent. As you can see, they are not independent, as many counts were above expected value, and one was significantly lower than expected value: Vision/Sound. Overall I was happy with the results of the project, which showed that different pairs of synesthetic words show up at dependent rates, resulting in certain pairs frequently appearing and others rarely showing up despite containing common single-sense words.

## 7. Experimental Evaluation

My dataset was a wide range of literature styles from classic English authors. It featured over one thousand works, split up into over ten thousand samples of text, from nearly three hundred authors. The data was split over several tables, to contain additional meta data such as book id, book name, and author name.

While the data was suitable for my purpose, the wide variation in the length of works was slightly vexing. Some works were hundreds of thousands of words, while others were very short, especially songs or poems. Had the dataset included genre, I may have been able to sort out those shorter pieces and focused on extended works of literature.

My performance metrics for this project were straightforward, since I was mainly interested in raw counts and ratios. My data was recorded in total number of instances of descriptive sentence fragments, to get an estimate for the prevalence of synesthetic types. The most interesting part of the analysis was probably the difference graph, which showed how combinations of senses can be very differently distributed than single senses, with certain combinations having higher or lower rates than might be expected.

For the sake of computation time, I adjusted my original plan from using spaCy pattern matcher to using simple word flagging. While using spaCy, my runtimes were long as it was running extra processes behind the scenes. When switching to more basic word matching, the runtime was decreased significantly. It took my script under 8 minutes to process and sort every sentence fragment in 10,000 text samples, using a single thread.

## 8. Future Work

The main upgrade to this project would be developing a sophisticated model for detecting synesthetic connections.  Right now, it’s built on matching keywords, which does an okay job, but also produces a fair amount of false negatives and false positives. English words can mean many different things, and limiting a word to just one category, while necessary to prevent multi-sensory detection from going off from a single word, also doesn’t capture the full breadth of meaning.  In essence, this project provides a solid estimation of synesthetic rates, and not definitive values.

In order to make a model sophisticated enough to detect nuance in meaning and word choice, a machine learning model would have to be trained on thousands of examples of synesthetic language.  Finding this training dataset would prove challenging, and tuning the model would also be a daunting task.  We are only now reaching a point where something like this would be feasible, and I dont have the resources or training set to create this model.

A more reasonable path for future work would be to branch out into different eras of literature, or different languages.  Seeing how other languages authors used synesthesia would be interesting, as would seeing how modern writers liked to use it.  I imagine it would show up more, since modern writing puts an emphasis on flowery, creative descriptions, as well as tends to use longer sentence fragments, but it’s hard to know without running the data.

## References

1. XC Zhang (Raynard). 2020. Classic English Literature Corpus and Meta Data. Retrieved 4/28/2024, from https://www.kaggle.com/

2. Banissy MJ, Jonas C, Cohen Kadosh R. Synesthesia: an introduction *(https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4265978/)*. *Front Psychol*. 2014;5:1414. Published 2014 Dec 15.

3. Zhao Q., Ahrens K., Huang C.  Linguistic synesthesia is metaphorical: a lexical-conceptual account *(https://doi.org/10.1515/cog-2021-0098).  Cognitive Linguistics.*  2022;553-583.  Published 2022, Jun 24.