An Evaluation of C.K. Moncrieff's Translation of Proust's *In Search of Lost Time* Using Statistics and Machine Learning

Introduction:

Last semester, I read *Swann's Way,* the first volume of Marcel Proust's momentous *In Search of Lost Time (ISOLT)* - and it left a lasting impression on me.

In Search of Lost Time holds the Guinness World Record for the longest novel ever written, and since its publication in 1913, has moved and inspired generations of readers and writers. Virginia Woolf famously quipped about it: "What remains to be written after that?"

Swann's Way greeted me to the wonderful world of language- not as a means to signify ideas and intentions, but as an aesthetic, or art. I vividly remember sitting in the Jonathan Edwards College library past midnight re-reading passages while listening to Arthur Rubinstein's recording of *Chopin Nocturnes*, tranquilized by the beauty of Proust's pen, the profundity of his revelations, and the humors of his characters.

Simultaneously, reading Proust also opened my eyes to the problem of translation: there was so much beauty in translation, but would the ethos of this masterwork affect me in the same way if I read the original French? What is lost - and gained- in translating ISOLT, a work that lets language speak for itself?

I chose to explore these questions with statistics, machine learning, and natural language processing (NLP), focusing on comparing the C.K. Scott Moncrieff translation and French text. I acknowledge that many aspects of ISOLT and literature as a whole cannot and should not be reduced to numbers and data. Hence, my findings are not meant to be taken as matter of fact; rather, they are an interesting window into the tangible differences between the quintessential translation of ISOLT and the original ISOLT, shedding insight into the limits and achievements of mapping objective mathematical metrics onto the literary arts.

In this project, I will first compare the word frequencies and sentence length distributions of the Moncrieff translation and original French text of *Swann's Way*.

Then, I will conduct sentiment analysis in three phases: I will 1) compare the sentiments of the Moncrieff to the original by section, 2) trace the sentiment arcs throughout the novel of both texts and 3) map out the semantics of each paragraph in both the translation and original relative to one another.

Finally, I will identify and compare the most prevalent topics in both the Moncrieff translation and the original French, evaluating how similar the extracted topics are between the two texts.

A Comparison of Word Frequencies and Sentence Length Distributions:

Methods:

My aim in this section was to compare the word frequencies and sentence lengths of both the French text and Moncrieff translation. I sought to answer: *does the Moncrieff translation maintain the French text's variety in vocabulary? How successfully does Moncrieff navigate syntactical challenges regarding translation when it comes to preserving Proust's stream-of-consciousness, long, winding sentences? When does Moncrieff preserve sentence structure- and when does he break apart a French sentence into multiple English ones?*

To separate the text into individual sentences, I first pre-processed the text to remove all special characters after common abbreviations such as "Mme", "Dr", etc. for both languages. Then, I used a sentencizer from spaCy, a NLP Python library, to count sentence length. I used spaCy's "en_core_web_sm" model, a small English model trained on written web text, to parse English sentences, and I used spaCy's "blank" French model, a minimal French model, to parse French sentences.

Word frequency analysis was conducted using the built-in Counter python function in conjunction with the spaCy models.

Results:

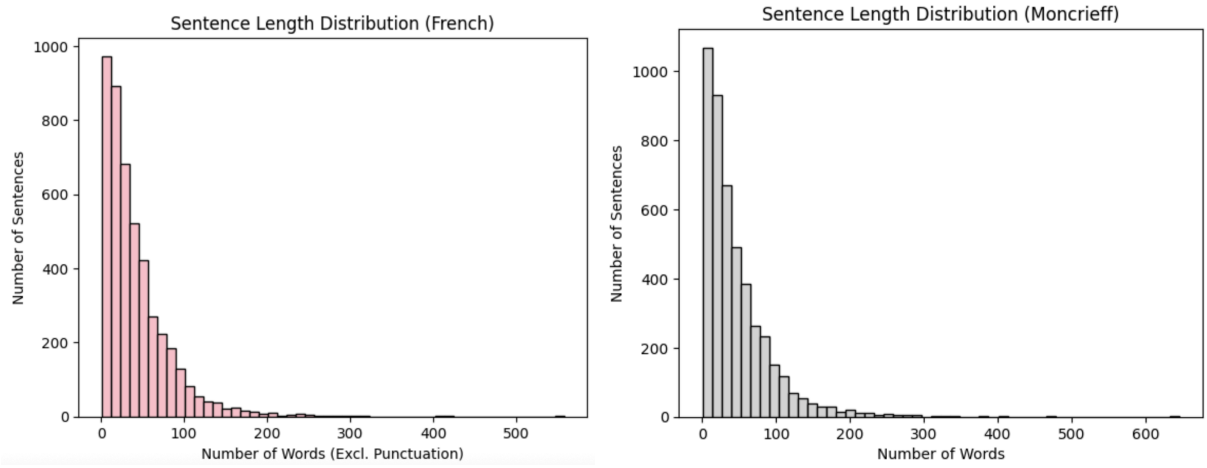

Figure A: Sentence length distribution histograms of French and Moncrieff text.

At a first glance, the sentence length distribution of the French text and the Moncrieff translation look nearly identical- very skewed right. However, the Moncrieff translation clearly has more sentences, with the leftmost bin containing over 1000 sentences, compared with the French text which has fewer than 1000 sentences. Furthermore, there are noticeable outliers in both the French text and Moncrieff translations on the right end. However, the Moncrieff edition's outlier sentences are longer.

As a result, it appears that Moncrieff broke up a lot of Proust's sentences, while maintaining Proust's sentences of extreme length.

A few questions arise- *Is Moncrieff's separation of Proust's medium-length sentences into shorter ones a conscious decision, a result of language syntax differences, or Moncrieff's preference?* Simultaneously, *why did Moncrieff choose to keep Proust's long sentences intact?*

The longest sentence is 646 words in the Moncrieff edition:

> But I had seen first one and then another of the rooms in which I had slept during my life, and in the end I would revisit them all in the long course of my waking dream: rooms in winter, where on going to bed I would at once bury my head in a nest, built up out of the most diverse materials, the corner of my pillow, the top of my blankets, a piece of a shawl, the edge of my bed, and a copy of an evening paper, all of which things I would contrive, with the infinite patience of birds building their nests, to cement into one whole; rooms where, in a keen frost, I would feel the satisfaction of being shut in from the outer world (like the sea-swallow which builds at the end of a dark tunnel and is kept warm by the surrounding earth), and where, the fire

keeping in all night, I would sleep wrapped up, as it were, in a great cloak of snug and savoury air, shot with the glow of the logs which would break out again in flame: in a sort of alcove without walls, a cave of warmth dug out of the heart of the room itself, a zone of heat whose boundaries were constantly shifting and altering in temperature as gusts of air ran across them to strike freshly upon my face, from the corners of the room, or from parts near the window or far from the fireplace which had therefore remained cold--or rooms in summer, where I would delight to feel myself a part of the warm evening, where the moonlight striking upon the half-opened shutters would throw down to the foot of my bed its enchanted ladder; where I would fall asleep, as it might be in the open air, like a titmouse which the breeze keeps poised in the focus of a sunbeam--or sometimes the Louis XVI room, so cheerful that I could never feel really unhappy, even on my first night in it: that room where the slender columns which lightly supported its ceiling would part, ever so gracefully, to indicate where the bed was and to keep it separate; sometimes again that little room with the high ceiling, hollowed in the form of a pyramid out of two separate storeys, and partly walled with mahogany, in which from the first moment my mind was drugged by the unfamiliar scent of flowering grasses, convinced of the hostility of the violet curtains and of the insolent indifference of a clock that chattered on at the top of its voice as though I were not there; while a strange and pitiless mirror with square feet, which stood across one corner of the room, cleared for itself a site I had not looked to find tenanted in the quiet surroundings of my normal field of vision: that room in which my mind, forcing itself for hours on end to leave its moorings, to elongate itself upwards so as to take on the exact shape of the room, and to reach to the summit of that monstrous funnel, had passed so many anxious nights while my body lay stretched out in bed, my eyes staring upwards, my ears straining, my nostrils sniffing uneasily, and my heart beating; until custom had changed the colour of the curtains, made the clock keep quiet, brought an expression of pity to the cruel, slanting face of the glass, disguised or even completely dispelled the scent of flowering grasses, and distinctly reduced the apparent loftiness of the ceiling.

This appears to be a direct translation of longest sentence in the French edition (537 words):

Mais j'avais revu tantôt l'une, tantôt l'autre, des chambres que j'avais habitées dans ma vie, et je finissais par me les rappeler toutes dans les longues rêveries qui suivaient mon réveil; chambres d'hiver où quand on est couché, on se blottit la tête dans un nid qu'on se tresse avec les choses les plus disparates: un coin de l'oreiller, le haut des couvertures, un bout de châle, le bord du lit, et un numéro des Débats roses, qu'on finit par cimenter ensemble selon la technique des oiseaux en s'y appuyant indéfiniment; où, par un temps glacial le plaisir qu'on goûte est de se sentir séparé du dehors (comme l'hirondelle de mer qui a son nid au fond d'un souterrain dans la chaleur de la terre), et où, le feu étant entretenu toute la nuit dans la cheminée, on dort dans un grand manteau d'air chaud et fumeux, traversé des lueurs des tisons qui se rallument, sorte d'impalpable alcôve, de chaude caverne creusée au sein de la chambre même, zone ardente et mobile en ses contours thermiques, aérée de souffles qui

nous rafraîchissent la figure et viennent des angles, des parties voisines de la fenêtre ou éloignées du foyer et qui se sont refroidies;--chambres d'été où l'on aime être uni à la nuit tiède, où le clair de lune appuyé aux volets entr'ouverts, jette jusqu'au pied du lit son échelle enchantée, où on dort presque en plein air, comme la mésange balancée par la brise à la pointe d'un rayon--; parfois la chambre Louis XVI, si gaie que même le premier soir je n'y avais pas été trop malheureux et où les colonnettes qui soutenaient légèrement le plafond s'écartaient avec tant de grâce pour montrer et réserver la place du lit; parfois au contraire celle, petite et si élevée de plafond, creusée en forme de pyramide dans la hauteur de deux étages et partiellement revêtue d'acajou, où dès la première seconde j'avais été intoxiqué moralement par l'odeur inconnue du vétiver, convaincu de l'hostilité des rideaux violets et de l'insolente indifférence de la pendule qui jacassait tout haut comme si je n'eusse pas été là;--où une étrange et impitoyable glace à pieds quadrangulaires, barrant obliquement un des angles de la pièce, se creusait à vif dans la douce plénitude de mon champ visuel accoutumé un emplacement qui n'y était pas prévu;--où ma pensée, s'efforçant pendant des heures de se disloquer, de s'étirer en hauteur pour prendre exactement la forme de la chambre et arriver à remplir jusqu'en haut son gigantesque entonnoir, avait souffert bien de dures nuits, tandis que j'étais étendu dans mon lit, les yeux levés, l'oreille anxieuse, la narine rétive, le cœur battant: jusqu'à ce que l'habitude eût changé la couleur des rideaux, fait taire la pendule, enseigné la pitié à la glace oblique et cruelle, dissimulé, sinon chassé complètement, l'odeur du vétiver et notablement diminué la hauteur apparente du plafond.

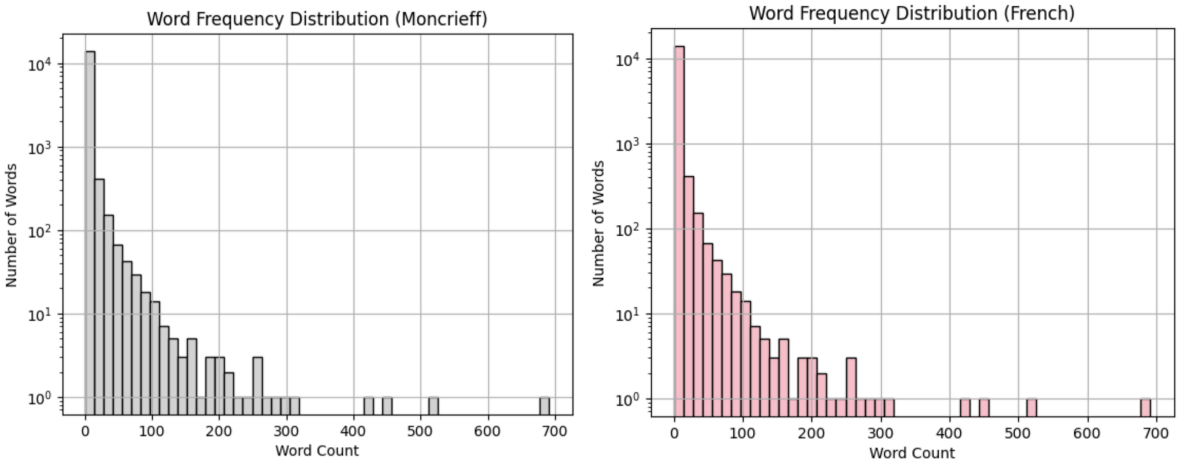

Figure B: Word Frequency Distributions of the Moncrieff Translation and French Text

The word frequency distributions of both the Moncrieff translation and French are identical. I found it astounding that 10,000 words are in the leftmost bin for the Moncrieff translation and French, highlighting the success of the Moncrieff translation in capturing the diverse vocabulary Proust employs in the French. "Swann" is the most used word, with 691 occurrences in the French text and 690 occurrences in the Moncrieff.

Finally, I explored whether the word frequency distributions of both texts follow **Zipf's Law- a law that states that the frequency of an observation is inversely proportional to its rank.** In natural language, the most frequent word tends to occur about twice as often as the second most frequent, three times as often as the third, and so on. When plotting the logarithm of word frequency against the logarithm of rank, the points should approximately fall along a straight line with a slope close to –1. A high R-squared value would indicate that the data closely follows this pattern.

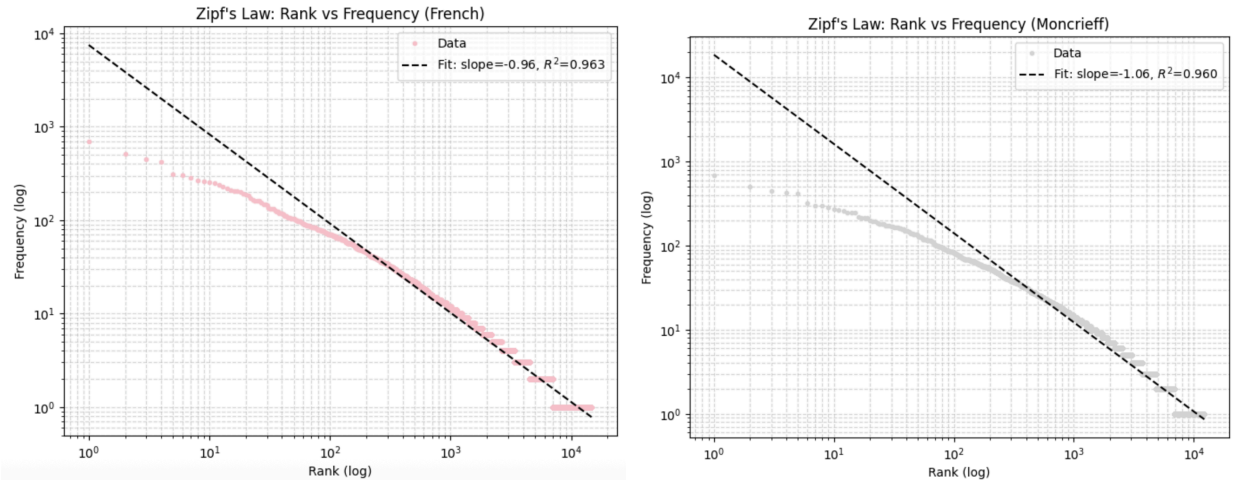

Figure C: Zipf's Law Graph of French Text and Moncrieff Translation. The dotted line represents what should be observed if the law was perfectly followed by the text.

The graphs for both texts show that the regression line has an approximate slope of -1, while the R-squared values of 0.963 and 0.960 show satisfactory fit of the data to the regression line. In other words, *Swann's Way* indeed closely follows Zipf's law- the frequencies of words are inversely proportional to their ranks.

To my mind, it is Proust's extraordinary breadth of vocabulary that endows Swann's Way with its power to leave lasting impressions. Diction plays a huge role in language; observing that the Moncrieff translation follows Proust's word variety highlights that Moncrieff valued diction as key to a successful translation. Simultaneously, the fact that both texts closely follow Zipf's Law- a law that most texts should follow- indicates that Proust's unique emotional impact is not solely projected through variety for the sake of variety. It's still the magic in the words that Proust uses, and the carefully selected words that Moncrieff chose to use in his translation, that make *Swann's Way* powerful.

Sentiment Analysis:

Methods:

I used a multilingual model from Hugging Face, an open source repository with deep learning natural language processing models, called “bert-basemultilingual-uncase-sentiment.” This is a fine-tuned model based on BERT, or the Bidirectional Encoder Representation from Transformers model introduced by researchers at Google in 2018.

What makes BERT unique is that it was created to capture an entire sentence’s sentiment; it **bidirectionally** looks at sentences from left to right and right to left all at once.

Upon being fed a sentence, first, BERT tokenizes a sentence’s words into words or subword chunks.

For example, if BERT takes in “The cat sat on the mat.” as an input, BERT would break up the sentence into: ['[CLS]', 'the', 'cat', 'sat', 'on', 'the', 'mat', '.', '[SEP]'], where CLS is a special token added to the beginning, and SEP is a special token that marks the end of a sentence. Then, BERT assigns an embedding to each token, or a vector that encapsulates the token based on the meaning of the token and the context of the token. Once each token is converted into a number, BERT passes the token embeddings through each of its encoder layers, which transform the embeddings and produces an output that aligns with the task the model has been assigned to do.

In the case of this specific model I used, the task of this multilingual BERT model is to rate sentences on a scale of 1 to 5, 5 being the most positive and 1 being the most negative.

Results:

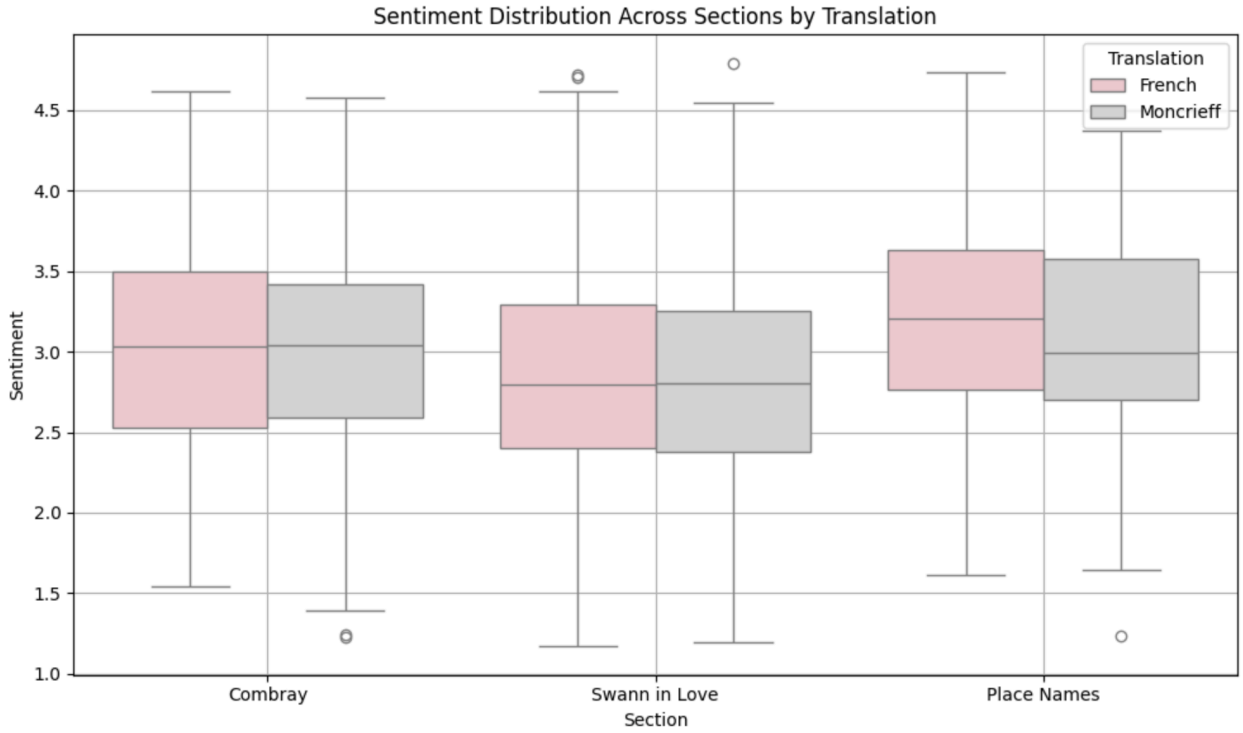

Figure D: Box Plots showing the sentiment distribution of paragraphs in each section of *Swann's Way*.

All in all, BERT's analysis suggests that Moncrieff's translation succeeded in preserving the sentiment of the original French text. It is interesting to note that for all sections and language of text, the sentiments of each paragraph center at around 3, which signifies a neutral sentiment. The biggest difference among translations occurs in Place Names, where the Moncrieff translation's median sentiment is noticeably lower than that of the French (2.9 compared to 3.3).

However, when looking at specific sentences, such as the famous first sentence: "Longtemps, je me suis couché de bonne heure", the text sentiments vary greatly.

```
text = "Longtemps, je me suis couché de bonne heure."
print("Sentiment score:", get_sentiment_score(text))
```

```
Sentiment score: 4.014577
```

```
text = "For a long time, I went to bed early."
print("Sentiment score:", get_sentiment_score(text))
```

```
Sentiment score: 2.974402
```

The Moncrieff is given a sentiment score of 2.97 (slightly below neutral), while the French is given a sentiment score of 4.01 (above neutral). I hypothesize that the word “bonne” makes the model interpret the French text as positive. Furthermore, because the model is mainly trained on product reviews, the words ‘long time’ are typically not used in a positive connotation in that context, which likely brings the Moncrieff translation to slightly negative.

Sentiment Arc:

Methods:

I was also curious about how the sentiment of the novel would shift over time. Thus, I recorded and graphed the sentiments of both texts as a function of their paragraph index, or their location in the text.

To smooth out the curve and remove noise from the data, I employed a rolling average- the model would calculate the average sentiment of not only the current paragraph, but also 5 paragraphs before it and 5 paragraphs after it.

Results:

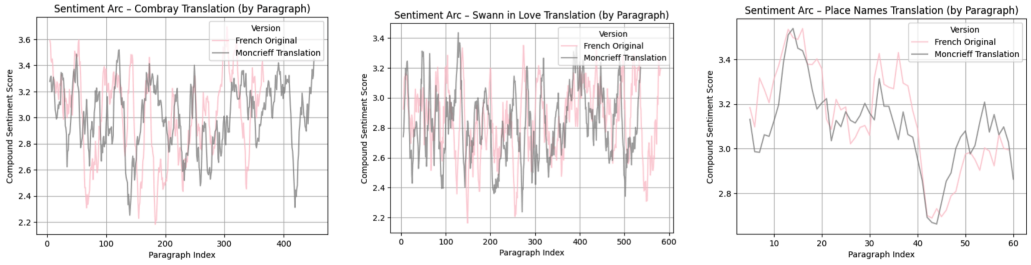

Figure E: Sentiment Arcs of all Three Sections of *Swann's Way* by paragraph index.

Interestingly, the number of paragraphs per section diverged more than I expected. For example, the Moncrieff translation of Combray has almost 100 more paragraphs than the original French, while the original French Swann in Love has around 50 more paragraphs than the Moncrieff translation. Nonetheless, it is clear that the semantic trajectories of all sections for each text are similar, highlighting that Moncrieff wonderfully captured the transitions between moods in the text.

Furthermore, the majority of the sentiments of the paragraphs are between 2.2 and 3.4, highlighting that there aren't major changes in the rolling average of the sentiment of the text, emphasizing that that the text's sentiment is not volatile.

I hypothesize that the lack of major volatility in both the translated version and the original French comes from Proust's neutral narrator. While the narrator often recalls both

positive and negative memories, the narrator relays his reactions to these memories and his emotions in those memories in a very objective manner.

Semantic Similarity:

Methods:

Furthermore, I explored how semantically similar paragraphs of each section and text were to one another. I used Language-Agnostic BERT Sentence Embedding (LaBSE), a BERT model presented in Google in 2020 that produces "language-agnostic cross-lingual sentence embeddings for 109 languages." LaBSE essentially creates a uniform scale of embeddings to capture passages of different languages, allowing for the comparison of text across multiple languages.

To visualize the relative distances of the embeddings, I used a t-SNE plot, or a t-distributed Stochastic Neighbor Embedding plot. The t-SNE takes in the embeddings, computes pairwise distances between the embeddings, and turns distances into probabilities that one passage would pick another as its neighbor.

Then, using Principal Component Analysis, the algorithm would initialize all embeddings into an x,y coordinate. Then, a new probability distribution corresponding to which data points are close to one another in the 2D space is generated, and through gradient descent, the 2D points are moved such that the distribution in the 2D space matches that of the original embedding space.

To interpret this plot, points closer to one another demonstrate higher similarity, and those further away demonstrate less similarity.

Results:

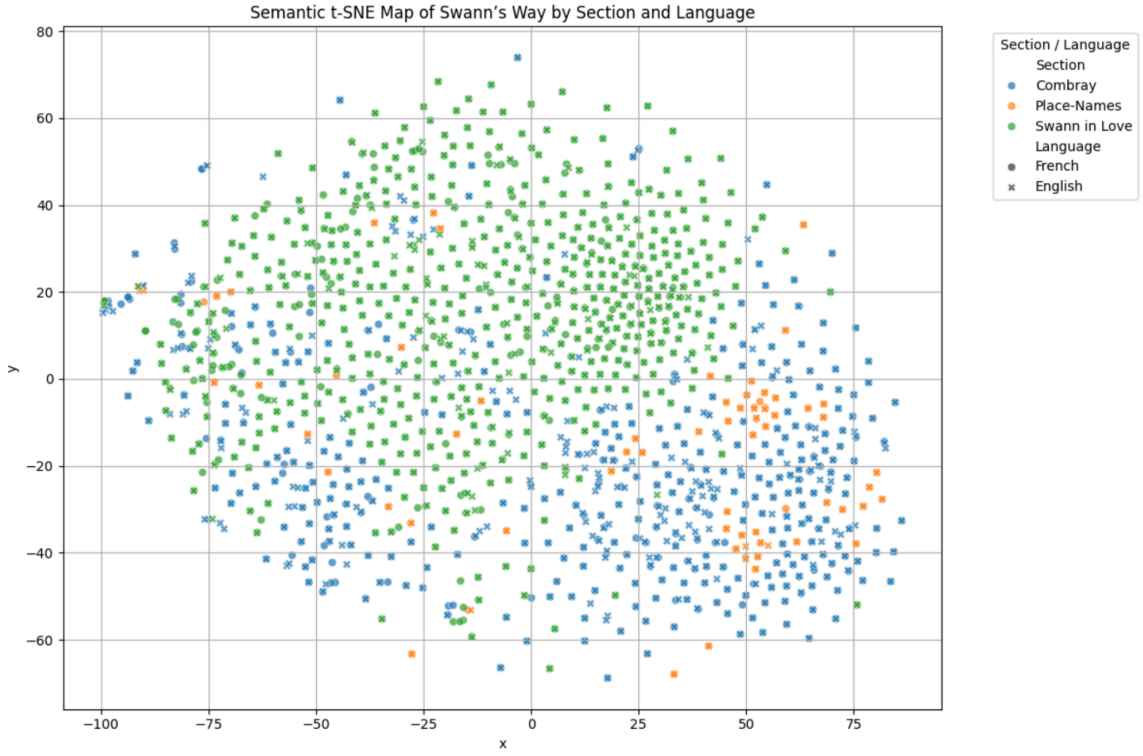

Figure F: T-SNE Map of the Sections of Swann's Way by language.

The t-SNE validates the Moncrieff translation's ability to capture the sentiment of the original text, with Xs (representing the translation) being very close to the dots (representing the French).

Furthermore, it is interesting to note that Swann in Love's paragraphs form a distinct cluster in the center of the plot. Meanwhile, Combray is more scattered, with Place-Names being the most scattered, with two small clusters formed around x = 50 in a sea of blue (Combray).

In my opinion, it makes sense that Swann in Love's paragraphs' semantics are closest to one another, as this section has the most well-defined plot out of all sections. While Combray is a scatter of spontaneous, involuntary memories, and Place-Names paints pictures of places that the narrator conjures up, Swann in Love follows episodes of Swann and Gilbertte's love story. Therefore, it makes sense that the paragraphs in Swann in Love would be more connected- and more semantically similar - to one another.

Topic Modeling:

Methods:

Finally, I explored what topics were most prevalent in the Moncrieff translation and original text. I employed BERTopic, a method of topic modeling. I used the spaCy French and English sentence transformers for the original text and Moncrieff translation, respectively. BERTopic converts sentences in French and English into vectors that capture the sentences' semantics and runs clustering algorithms to form distinct groups of sentences. Each distinct group is then considered to be a topic. To identify the topic, the model outputs the words that are most distinctive within the cluster to define the topic.

Specifically, I asked the model to generate 5 topics from both the French text and Moncrieff translation, with 1 topic serving as an 'outlier' topic category. To improve interpretability, I removed stop words and character names from being included in the topics.

Results:

Importantly, the Moncrieff and the French topic models both assigned almost half of the sentences to the outlier category. In the French model, 2435 out of 5621 sentences were assigned to the outlier topic, and 2525 out of 5082 sentences were assigned to the outlier topic.

Topic Number	Words in Moncrieff Topic Model	Words in French Topic Model
0	"Aunt, lover, thing, friend, little, eye"	"Être, faire, bien, avoir, grand, venir, aller, savoir, dire, voir"
1	"Day, paris, evening, dinner, night, carriage, door, home, time"	"Fleur, petit, voir, rose, air, eau, soleil, grand, arbre"
2	"Church, flower, sky, water, garden, way, tree, Combray"	"Piano, pianiste, bergotte, musique, sonate, musical, violon"
3	"Money, credit, diamond, wealth, creditor, barter"	"Artiste, oeuvre, photographie, beauté, art"

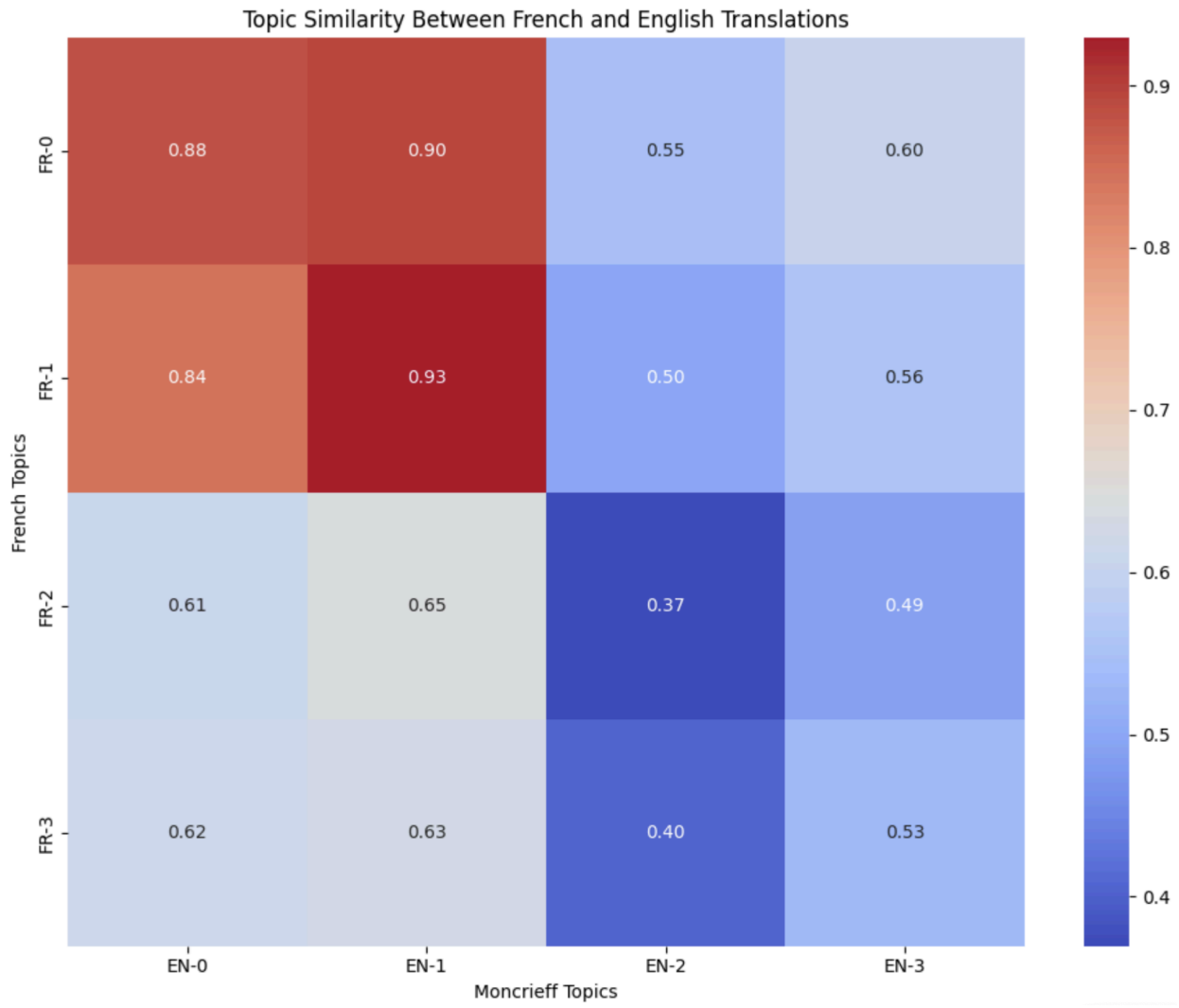

Figure G: A heatmap comparing topic similarity of Moncrieff and French text. Higher similarity scores represent topics that are more similar to one another.

While some topics (0 and 1) tend to not be interpretable, I find it interesting that the Moncrieff model does specifically identify a topic for nature (Topic 2) and another for money (Topic 3). Meanwhile, the French model identifies music (Topic 2) and art (Topic 3) as topics.

The heatmap above visualizes the drastic shift in the topics. While topics 0 and 1 are similar with high similarity values, the more interpretable topics (2 and 3) are very different between the French and English model.

In particular, the Moncrieff translation's topics appear to be more tangible, while the French text appears to focus more on aesthetics. *Could the divergence in translation be caused by differences in language? Or would the divergence in topics be a result of specific*

translation decisions made by Moncrieff, which removed many words to capture aesthetics and music?

Conclusion

In this project, I compared C.K. Scott Moncrieff's translation of Proust's *Swann's Way* to the original French text using statistical and natural language processing methods. In particular, I compared sentence length, word frequency, overall sentiment, sentiment arcs by section, semantic similarity, and topic modeling results of both texts.

I found that Moncrieff preserved Proust's longer sentences while separating a lot of Proust's shorter and medium-length sentences into shorter sentences. Furthermore, Moncrieff succeeded in preserving Proust's vocabulary variety, overall sentiment by section, and sentiment arcs.

By section, Swann in Love had the most uniform paragraphs in terms of semantic similarity, while Combray's paragraphs were the most scattered. Topic modeling results also varied between the two texts- in particular, for both texts, while the two most populated topics are uninterpretable, the latter two topics for the French text revolved around aesthetics (art, music), while the latter two topics for the English text revolved around tangible items (money, nature).

Ultimately, many of the revelations in this project (especially as it pertains to natural language processing modeling with transformers), with the exception of word frequency and sentence length, should be interpreted as indications about the efficacy of using objective metrics like statistics to measure literature, an art. Given that the sentence transformers and models were not trained specifically on Proust or literature in general and that sentiment and semantics in literature are subjective, these metrics alone should not dictate the validity and success of Moncrieff's translation.

Rather, this project is an enlightening exploration and attempt at evaluating the physical properties of the text. The findings in this project, for me, aim to get to the heart of what makes Proust's prose and language so evocative.