Slide 2: What are we after, and what to look out for.

- "Researchers interested in investigating the causes behind the variations in the styles produced by different translators focused on the literary side of the analysis."
- "On the other hand, researchers who focused on identifying measurable features of stylometric identification focused on the computational linguistic side of the analysis."
- Abridgement is not a random process; editors often cut specific types of content (e.g., descriptive passages), which can fundamentally alter our results.
- "Multiple studies have shown that when using stylometric analysis, translated texts often cluster by their original author rather than by their translator."
- Statistical techniques are more stable on larger datasets.

Slide 3: Overview of Stylometric Approaches

- **Classic Stylometric Tools:**
  - **Most Frequent Words (MFW)**These are powerful because their use is largely unconscious and topic-independent, making them good indicators of a translator's habits."
  - "K-index is a measure that was proposed by Yule (1944). The measure monotonically increases as the high-frequency words in the text increases."
  - "W-index is originally proposed by Brunet (1978), who suggested that this measure is not affected by text length, and it is author specific. W-index increases when the number of different words increases."
  - **N-gram Frequencies**, which look at sequences of words or characters to capture habitual phrasing and morphological tendencies.
- **Novel Stylometric Approaches (Network Stylometry):**
  - Addressed parsing the translator from the original author.
  - Still vulnerable to text length.

Slide 4: Methodological Considerations (Stop Words & Data Sets)

- **On Stop Words:**
  - "Interestingly, our approach to stop words will differ depending on the method. For the **novel network analyses**, we will *keep* the stop words. Research shows that these common function words are essential to the structural patterns that network motifs capture and removing them can actually reduce the accuracy of the analysis."
  - "However, for the **classic approaches** that rely on lexical frequency, I recommend we *remove* stop words after we've handled the text length issue. This will help focus the analysis on more distinctive word choices."

Slide 5: Strategy for Handling Abridged Texts

There are different matrices we can build, and I recommend a directed, weighted word adjacency matrix. This method is specifically designed to capture the local syntactic structure and word-order patterns that represent a translator's unique style. It's not just about *which* words are used, but *how* they are strung together. This is what the Abbass paper found to be a more reliable "thumb-print" than simple word frequencies

(Co-occurrence Matrix: Connects words that appear in the same sentence or paragraph. This is better for thematic analysis (which words are talked about together) but it **discards the syntactic word-order information** that is essential for analyzing a translator's structural style.)

- **1. Split the Corpus:**
- **2. Normalize for Length Using 'Chunking':**
  - Could we use chapters as a more organized form of chunking?
- **3. Rank-Based Features:**
  - "We can use features that are naturally more resistant to changes in text length. Instead of comparing raw frequency counts, we can compare the **rank order** of the most important words, such as those identified by centrality measures. The core structural words of a text often remain central even in a shortened version."

Slide 6: Advanced Visualization: Spectral Embedding

- **What it is:** "Spectral Embedding is an unsupervised machine learning technique that creates a simple 2D or 3D 'map' of our texts. On this map, texts with similar styles will appear clustered closely together."
- **How it works:**
  - "The process begins by building a similarity graph that connects all five texts, with the connections weighted by stylistic similarity."
  - "It then uses a mathematical tool called the **Graph Laplacian** to analyze the fundamental structure of this network."
  - "By calculating the eigenvectors of this Laplacian matrix, the algorithm can map the complex relationships onto a simple scatterplot."

- **Why it's better than PCA for this project:** "Spectral Embedding is specifically designed to preserve the *local network structure*, meaning it does an excellent job of showing which texts are most similar to each other. This is a significant advantage for a small corpus like ours, where every relationship is important."