From Human Strategy to Algorithmic Bias: A Quantitative Framework for Translation Analysis

Modern Artificial Intelligence (AI) systems allow anyone to generate translations with unprecedented ease. The widespread availability of these tools raises questions about the characteristics of machine-generated texts, especially for works of high cultural and interpretive value. In the past, readers relied on institutional translations aligned with a specific interpretive tradition; today, any user can generate a new translation. This creates uncertainty about a model's *objectivity*: which tradition will it follow, and will it do so consistently?

The study of translation is approached by two fields: Computer Science (CS) and Translation Studies (TS). Research in CS focuses on metrics that evaluate style *correctness* rather than *describing* it. TS, in turn, offers theoretical frameworks for such characterization, viewing every translation as an adaptation for a specific purpose and audience. However, its quantitative methods require large text collections for statistical analysis of a translator's decisions, precluding sentence-level examination. Furthermore, foundational theories in TS, like the concept of a translation strategy, were developed assuming a human translator making conscious and consistent decisions. In AI, a model's output stems not from deliberate choices but from the statistical distribution of its training data and an element of randomness in word selection. What we interpret as *strategy* in a human is more akin to a model's learned *bias*. There is thus a clear need for methods that describe, rather than evaluate, this bias. These methods should enable stylistic analysis at the sentence level while being scalable enough to examine the consistency of this bias across entire works.

To address this need, we propose two new computational methods for characterizing stylistic and semantic tendencies in machine translations. To validate them, we will build a new, multilingual corpus of New Testament translations – a text with the world's richest translation history, whose many versions often explicitly indicate a denominational tradition or target audience. Its high interpretive density means that the choice of a single word can reflect a particular theological stance, making it ideal for studying bias in a broad context. We will collect over 600 texts from a dozen websites in five languages: English, Spanish, French, Polish, and Italian. These texts will be enriched with metadata, such as denominational tradition and publication year, allowing for comparative analysis over time, across traditions, and between different languages.

Once the corpus is constructed, we will validate our first method, which uses interlinear translation as a computational baseline for characterizing translation style. An interlinear text is a word-for-word translation that preserves the source text's structure while using the target language's words. Our hypothesis is that the difference between the vector representations of the interlinear text and the translation under study (which we call the *intervention vector*) represents the stylistic and semantic choices made by the translator. We will test this by examining whether clustering algorithms group these vectors according to known metadata (religious tradition, century of origin) and by decoding the vectors themselves to translate mathematical data back into descriptive terms and identify specific stylistic tendencies, such as archaization.

Using our second method, we will investigate iterative round-trip translation as a technique for amplifying a model's latent biases. Our hypothesis is that cyclically translating a text from language A to B and then from B back to A acts as a feedback loop. Each translation cycle can subtly reinforce the model's stylistic preferences, making them more pronounced and easier to observe. We will conduct two types of experiments: one where we start this process from the original Greek text, and another where we start with existing human translations to test whether the model's biases can be guided toward a specific tradition.

As a result, the project will deliver two methods for measuring the characteristics of translations and a multilingual corpus to support this research. Our goal is to bridge the methodological gap between Computer Science and Translation Studies by creating new ways to investigate translation at a scale that was previously difficult to achieve. Ultimately, this will contribute to a better understanding of the AI technologies that a growing number of people use daily. Although the project focuses on machine translation, we hope its results will also contribute to the study of human translation.