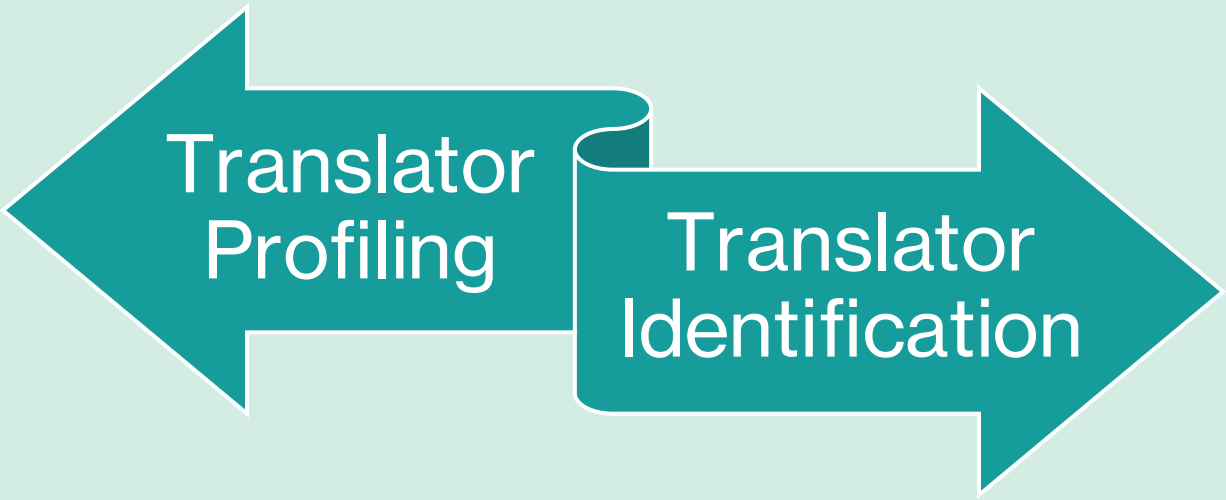
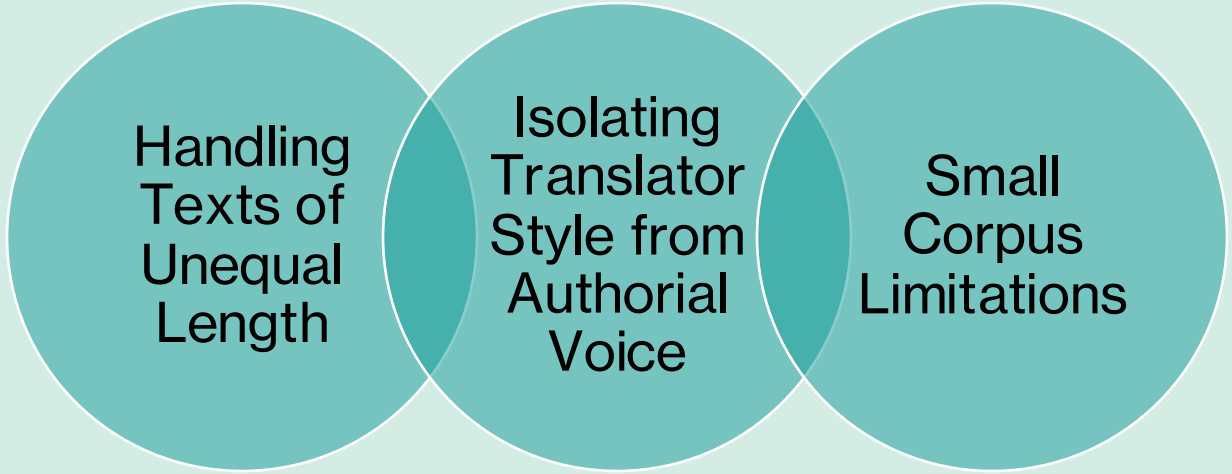
Stylometric Analysis

Classic Techniques and Novel Approaches

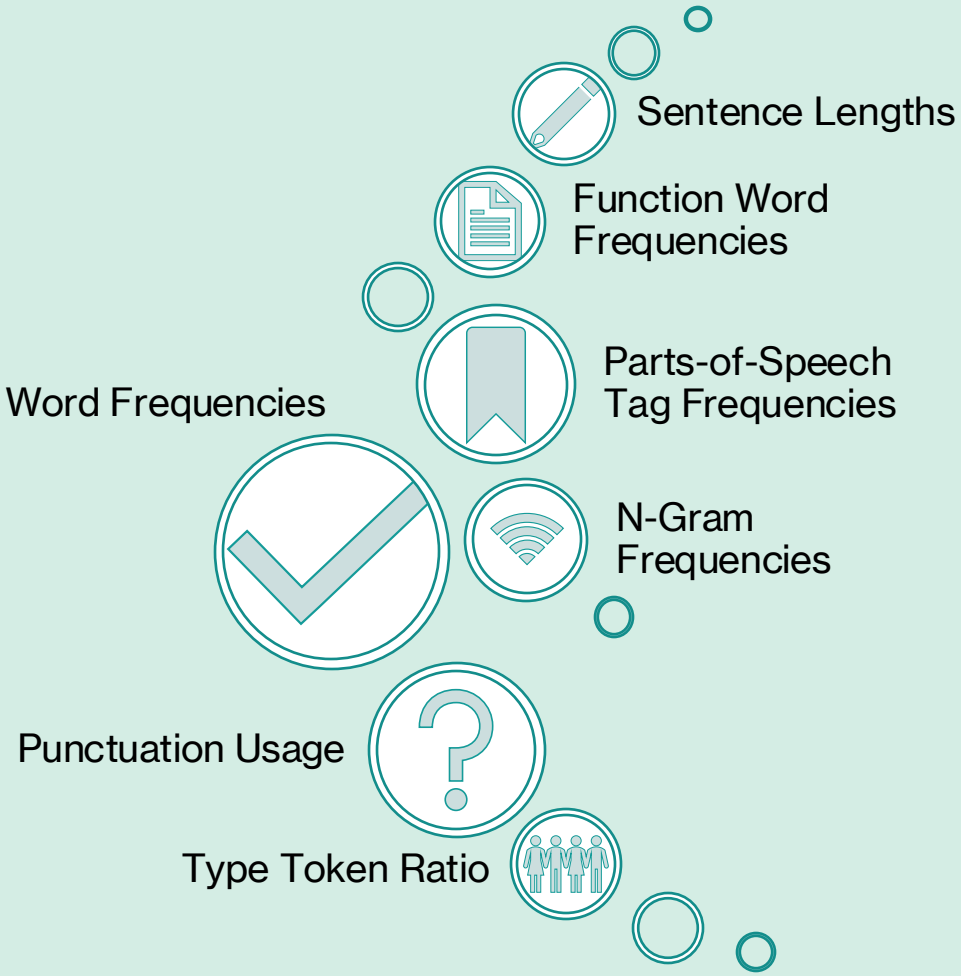
Objective



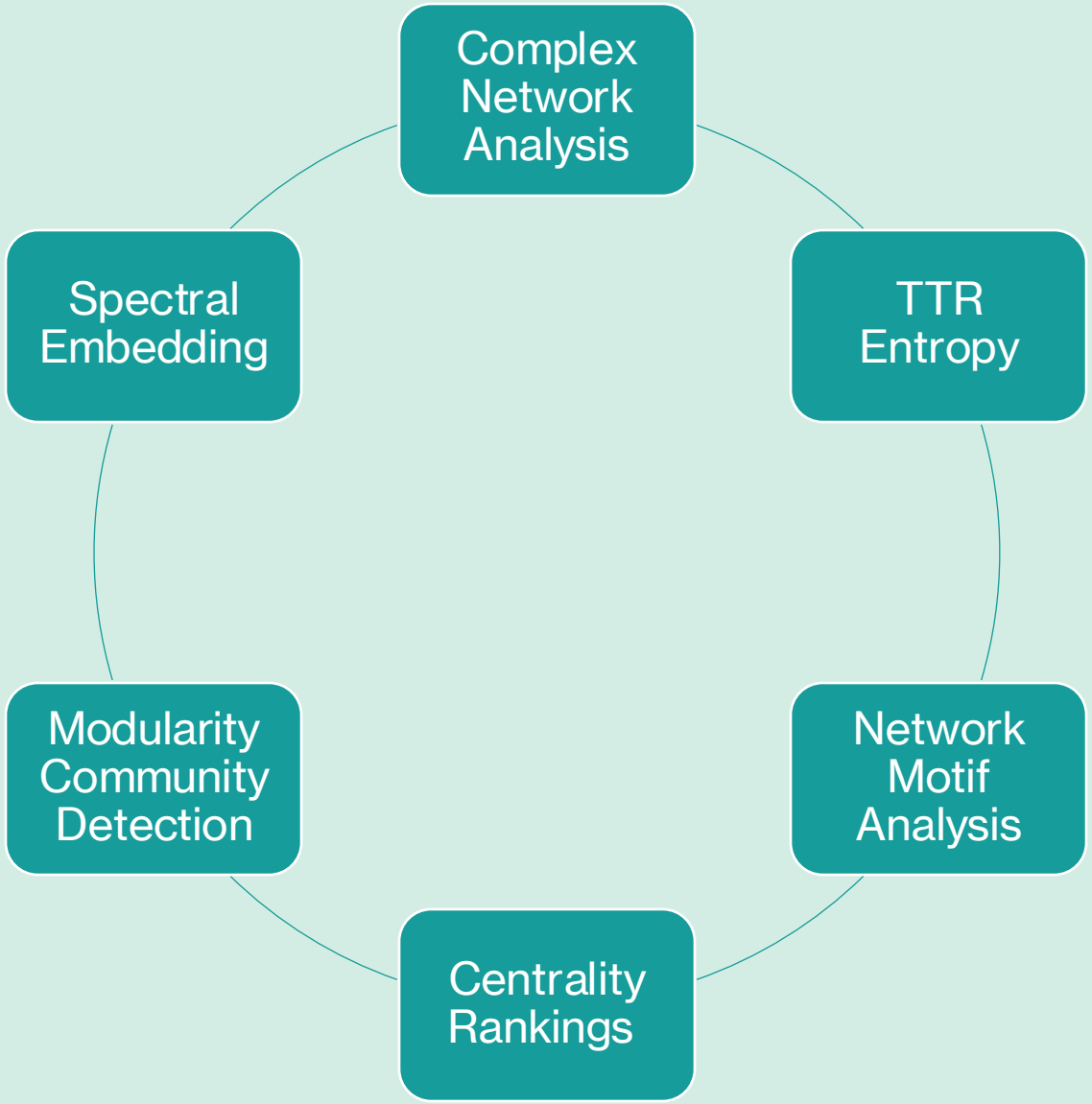
Challenges



Classic Analysis Techniques

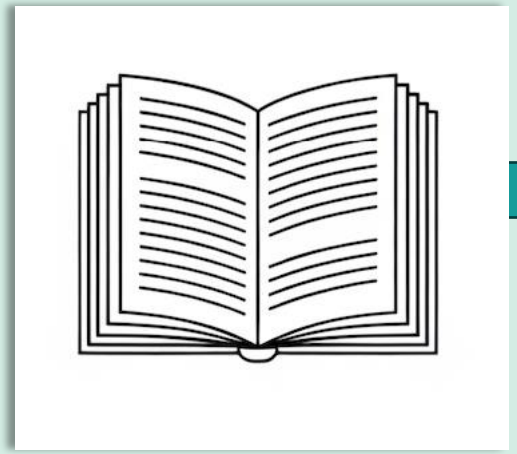


Novel Approaches



Complex Network Analysis

Step 1
Source Text



Step 2
Extraction

знать

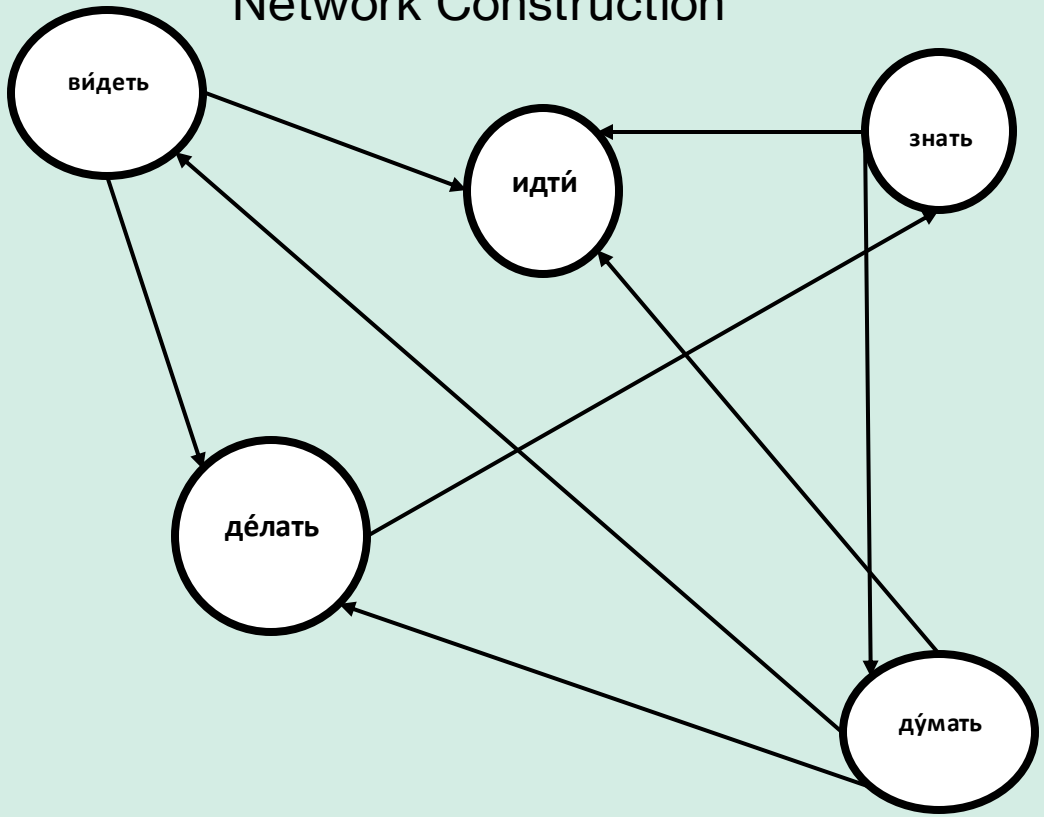
идти́

де́лать

ви́деть

ду́мать

Step 3
Network Construction



Bigram Adjacency Matrix

	знать	идти́	де́лать	ви́деть	ду́мать
знать	0	3	1	1	2
идти́	3	0	2	6	12
де́лать	1	2	0	0	1
ви́деть	1	6	0	0	3
ду́мать	2	12	1	3	0

Network Motif Extraction Counts vs Ranked Scaling

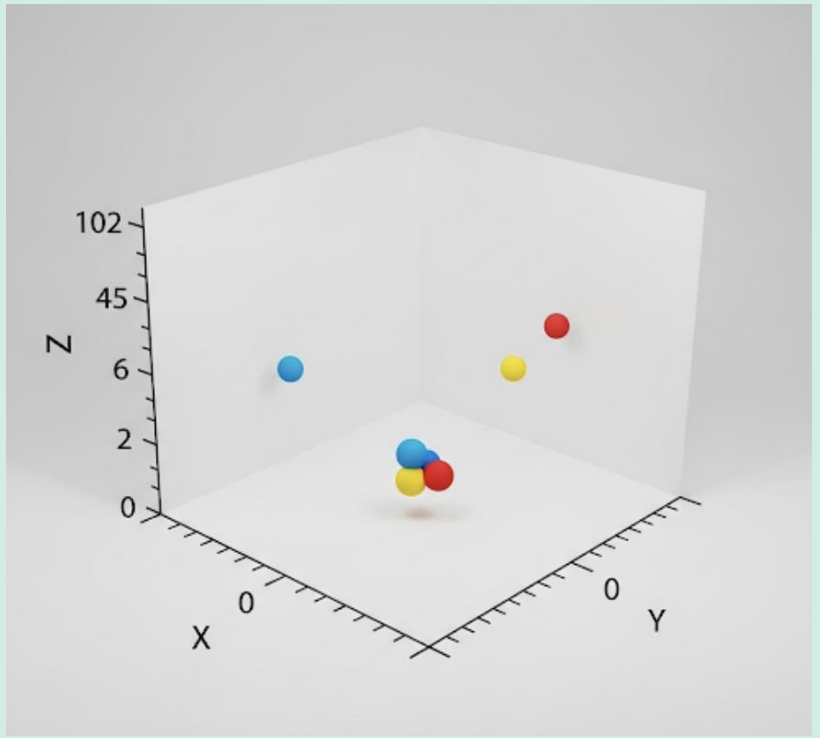
Translator	Motif Type A (raw count)	Motif Type B (raw count)	Motif Type C (raw count)
Translator 1 (Full)	150	25	80
Translator 2 (Full)	120	40	60
Translator 3 (Abridged)	75	10	50

Translator	Motif Type A (rank)	Motif Type B (rank)	Motif Type C (rank)
Translator 1 (Full)	1	2	1
Translator 2 (Full)	2	1	2
Translator 3 (Abridged)	3	3	3

Spectral Embedding

Cosine Similarity Matrix (DTM) without Stopwords:

	Ru1.txt	Ru2.txt	Ru3.txt	Ru4.txt	Ru5.txt
Ru1.txt	1.0000000	0.9752601	0.9637571	0.9698395	0.9647612
Ru2.txt	0.9752601	1.0000000	0.9862798	0.9898919	0.9859838
Ru3.txt	0.9637571	0.9862798	1.0000000	0.9909761	0.9934376
Ru4.txt	0.9698395	0.9898919	0.9909761	1.0000000	0.9983270
Ru5.txt	0.9647612	0.9859838	0.9934376	0.9983270	1.0000000



L = D − W

L = Laplacian Matrix

D = Diagonal Degree matrix . Each diagonal entry D_{ii} is the sum of weights of the edges connected to node i.

W = weighted adjacency matrix, where W_{ij} represents the similarity or weight between nodes i and j.

Recommendations

- **Use Spectral Embedding for Visualization:** This advanced technique is superior to PCA for our small corpus. It will create a more accurate map of the translators' stylistic clusters by better preserving local network structures.

- **Normalize for Text Length to Address Abridgement:** Use "chunking" or split the corpus before applying classic tools. This is critical because these tools are highly sensitive to text length and will produce invalid results otherwise.

- **Employ Rank-Based Scaling for Network Analysis:** For novel methods like network motifs and centrality, we will use ranks instead of raw counts. This minimizes length bias and is proven to be more effective at capturing the subtle, structural "thumb-print" of a translator's style.

Citations

- Cooper, D. L., Ogoltsev, D., & Ondrejcek, M. (2023). Type-token ratio and entropy as measures to characterize a forgery of oral-formulaic epics. *Oral Tradition*, *36*(1), 37–62.

- Eder, M. (2017). Visualization in stylometry: Cluster analysis using networks. *Digital Scholarship in the Humanities*, *32*(1), 50–64. https://doi.org/10.1093/llc/fqv061

- El-Fiqi, H., Petraki, E., & Abbass, H. A. (2019). Network motifs for translator stylometry identification. *PLoS One*, *14*(2), e0211809. https://doi.org/10.1371/journal.pone.0211809

- Zenker, F., & Kyle, K. (2021). Investigating minimum text lengths for lexical diversity indices. *Assessing Writing*, *47*(2), 100505. https://doi.org/10.1016/j.asw.2020.100505