Thesis Proposal

Title

Trend Estimation of Semantic Change in Bangla: A Computational Approach Using N-grams and Word Embeddings

---

Problem Statement

Language evolves constantly. Words often change the way they are used over decades, reflecting social, cultural, and technological shifts. In Bangla, for example, English loanwords like "কুল" (cool) have shifted from meaning lineage/class to meaning "trendy" or "stylish" in modern usage. Similarly, words like “ডিজিটাল” barely existed before 1990 but now permeate everyday speech.

This phenomenon is **semantic change**: the trend over time in how a word’s *meaning* or *usage context* evolves.

Our goal is to **estimate these trends** in Bangla language use over time, specifically:

- Detect which words have changed in meaning or usage
- Quantify how their contexts and associated meanings have changed
- Understand when these changes occurred (decade-wise or era-wise)

Such trend estimation is valuable for:

- Historical linguistics
- Building better Bangla NLP models
- Understanding cultural shifts encoded in language

However, **no large-scale computational work exists for Bangla** in this area. Most existing research is for English and other high-resource languages.

Research Objectives

1. Develop a computational method to estimate trends in semantic change for Bangla words over time.
2. Analyze large diachronic Bangla corpora from ~1950 to present.
3. Identify and explain specific words that have undergone semantic shifts.
4. Build case studies showing the evolution of usage trends.
5. Produce tools or resources (e.g. N-gram viewer) to visualize these trends.

Scope and Limitations

- Focus on Bangla only (no cross-linguistic comparison)
- Time range: ~1950–2025 (depending on corpus availability)
- Both formal (literature, news) and informal (web, social media) text
- Limitation: Scarcity of digitized, high-quality older corpora; requires OCR, cleaning, careful preprocessing
- Limitation: No existing "gold standard" labels for Bangla semantic change

---

Methodology

1. Corpus Building

- Gather historical Bangla text from multiple sources:
  • Literature (Wikisource, Vacaspati Corpus)
  • Newspapers (digitized archives)
  • Online news portals (Prothom Alo, Anandabazar)
  • Web/social media (KUMono, BanglaBERT data)
- Split into time periods (1950–1970, 1970–1990, 1990–2010, 2010–2025)
- Clean, normalize script, remove OCR errors

2. Trend Estimation via N-gram Analysis

- Count word-frequency trends
- Extract top collocates per period
- Compare collocation shifts
- Build an interactive N-gram Trend Viewer

3. Trend Estimation via Word Embeddings

- *Static Embeddings:*
  • Train Word2Vec/fastText per period
  • Align via Orthogonal Procrustes
  • Measure cosine-distance shifts & neighbor changes
- *Contextual Embeddings:*
  • Use BanglaBERT/mBERT to extract context vectors
  • Cluster instances, measure inter-period distribution distance

4. Case Studies

- Select words like কুল (cool), ভাইরাল (viral), ডিজিটাল (digital), স্বাধীনতা (independence)
- Present frequency graphs, collocates, embedding neighborhoods

5. Evaluation

- Expert review & dictionary comparison
- Native-speaker survey on perceived meaning shifts
- Correlate changes with historical events (e.g. 1971 war)

Proposed Workflow

Data Collection → Cleaning & Preprocessing → Time-Sliced Corpus

→ N-gram Analysis → Embedding Training & Alignment

→ Change Scoring → Contextual Analysis → Case Studies → Reporting

---

Proposed Algorithms and Tools

- **N-gram Analysis:** frequency counts, collocates, visualization
- **Static Embeddings:** Word2Vec/fastText, Procrustes alignment, cosine shifts
- **Contextual Embeddings:** BanglaBERT, clustering, distributional metrics
- **Visualization:** Matplotlib/Seaborn plots, interactive viewer prototype

---

References

- Hamilton et al. (2016). *Diachronic Word Embeddings Reveal Statistical Laws of Semantic Change*. ACL. [https://aclanthology.org/P16-1141.pdf]
- Truică et al. (2023). *Semantic Change Detection for the Romanian Language*. arXiv. [https://arxiv.org/abs/2305.10354]
- Kutuzov et al. (2018). *Diachronic Word Embeddings and Semantic Shifts: A Survey*. [https://arxiv.org/abs/1806.03537]
- Chaudhury et al. (2023). *VACASPATI: A Diverse Corpus of Bangla Literature*. arXiv. [https://arxiv.org/abs/2311.16545]
- Akther et al. (2022). *KUMono 350M Word Bangla Corpus*. IEEE Access. [https://ieeexplore.ieee.org/document/9926052]

Conclusion

This project will deliver a comprehensive computational approach for **trend estimation of semantic change** in Bangla. By combining n-gram statistics with modern embedding models, we will reveal how Bangla words have shifted in meaning and usage over decades. The outcome will be academically valuable and culturally insightful, pioneering future research in Bangla NLP and diachronic linguistics.