ChatGPT

Detecting Semantic Change in Bengali Over Time: Feasibility and Approaches

Background and Motivation

Language is never static – the meanings and uses of words shift over decades due to cultural, technological, and social changes. In English, for example, the word *"cool"* once primarily described temperature but later gained a popular meaning of *"fashionable" or "good."* This thesis project aims to track such **semantic change** within the Bengali (Bangla) language from around the 1950s to the present. Unlike cross-linguistic studies, the focus here is **diachronic**, examining how Bengali words' meanings have drifted within Bengali itself over time. Detecting these changes can shed light on linguistic evolution and is valuable for historical linguistics, lexicography, and improving NLP systems (which often assume static word meanings).

Bengali presents both an opportunity and a challenge for semantic change research. It is the world's seventh most spoken language, but it is relatively **under-resourced** in NLP research [1] [2]. Prior computational studies of semantic change have mostly focused on high-resource languages (English, German, etc.), so a Bengali study could fill a research gap. The project envisions using **n-gram statistics** alongside modern NLP models (like word embeddings) to quantify meaning shifts. By using a combination of **structured corpora** (e.g. literature, newspapers) and **unstructured text** (e.g. social media posts) from different eras, we can build a timeline of word usage. For instance, we might show how a Bengali word for "smart" or "cool" in the 1960s had only a literal meaning, but by 2025 it gained a new figurative meaning among younger speakers. Such case studies not only illustrate the phenomena but also make the research tangible. Overall, the motivation is twofold: to contribute to linguistic knowledge of Bengali and to develop techniques for **tracking semantic drift** in a low-resource setting.

Prior Work on Lexical Semantic Change Detection

Research on **lexical semantic change (LSC)** detection has grown in recent years, although Bengali-specific studies are scarce. In general, computational approaches model how a word's **contextual usage** changes across different time periods. A seminal work by Hamilton et al. (2016) introduced a framework using **distributional word embeddings** to quantify semantic change [3]. By training word vectors on historical text corpora and comparing them over time, Hamilton et al. uncovered statistical laws of semantic evolution – for example, *rarer words change meaning more rapidly, and polysemous words (with multiple senses) tend to have higher rates of semantic change* [3]. These findings highlight that frequency and initial breadth of meaning can influence how a word's sense evolves.

Most prior studies have centered on well-resourced languages due to data availability. For example, the **SemEval-2020 Task 1** provided benchmarks for semantic change in English, German, Latin, and Swedish, spurring many method developments. Approaches in those studies range from counting changes in word context distributions to training separate embedding models per era. In contrast, for Bengali (a low-resource language in this domain), published research is very limited. To our knowledge, there have not yet been large-scale published studies specifically on Bengali semantic shift using modern techniques – which

is precisely why this thesis could break new ground. There are some related efforts: for instance, a recent study examined **Bengali neologisms** that emerged during the COVID-19 pandemic [4]. That work used a semi-automated approach to identify newly coined or repurposed words in pandemic-related news, essentially tracking vocabulary changes in real time [4]. While neologism detection is a narrower task (focused on new words rather than changing meaning of existing words), it demonstrates growing interest in how Bengali evolves with events. Additionally, a few linguistics papers have looked at meaning change in specific contexts (e.g., Arabic loanwords in Bengali shifting meaning over centuries), but these were manual analyses [5].

Encouragingly, research on **low-resource languages** like Romanian has shown that semantic change detection methods can be successfully adapted beyond English. A 2023 study on Romanian compared static embeddings (Word2Vec) and contextual embeddings (ELMo) to detect meaning shifts [6]. They first validated their approach on English (using the SemEval-2020 dataset) and then applied it to Romanian, highlighting phenomena like words gaining or losing senses [7]. Their results underscored that the choice of model and the distance metric have a big impact on performance [7]. Similarly, other recent works explored **contextualized models** (e.g. BERT-based) for semantic change in languages such as Italian, Latin, or Swedish, finding that these models can capture subtle shifts and often outperform simple frequency-based measures [8] [9]. All this prior work provides a methodology foundation and suggests that doing a Bengali study is feasible – the main hurdle being the availability of suitable diachronic data and some necessary adaptations for the specifics of Bengali.

Bengali Diachronic Corpora: Available Resources (1950s–Present)

A critical step is obtaining **diachronic corpora** – text collections from different time periods – for Bengali. We ideally want data from around the 1950s onward, split into time spans (e.g., by decade or by specific eras). This will allow training separate models or computing separate statistics for each period and then comparing them.

1. Literary Texts: Literature is a rich resource for historical language. Many Bengali literary works (novels, plays, essays, poetry) from the 19th and 20th centuries are available in digital form. For example, the recently released **VACASPATI** corpus is a large collection of classical Bangla literature spanning several centuries [10]. It contains about 115 million words from works ranging from the 14th century up to the 21st century, with diversity in authors, genres, and eras [10]. Such a corpus is extremely valuable: one can filter subsets of VACASPATI by publication date (e.g., novels from the 1950s–60s vs. 2000s) to compare language usage. Literature from the 1950s in East Pakistan (now Bangladesh) and West Bengal might reflect the language of that time, albeit in a relatively formal register. Another source is **Bengali Wikisource**, where volunteers have digitized many public-domain books (including works by Tagore, Bankim Chandra, etc., mostly pre-1950 but also mid-20th century texts). These literary corpora offer relatively clean, structured text and can be segmented by time period (by author birth date or publication date) as proxies for earlier decades. However, one must be cautious: written literary Bengali, especially before the 1970s, often used the **Shadhu-bhasha** (formal register) which has since largely given way to **Cholit-bhasha** (colloquial register) in writing. Some words or grammatical forms common in older literature may have fallen out of use, which is itself a kind of linguistic change but more stylistic than semantic.

2. Newspapers and Magazines: Newspapers are ideal for capturing contemporary vocabulary and usage of their time. For Bengali, major newspapers started in the mid-20th century. For example, *Dainik Ittefaq* began in 1953 in Dhaka, and *Anandabazar Patrika* has been a leading Bengali daily in Kolkata for a century.

Digitized archives of such newspapers would be a goldmine of dated language data. In practice, some archives exist (e.g., libraries have microfilms or PDFs of old issues), but OCR (optical character recognition) may be needed to convert them into text. There has been work on compiling Bengali news corpora; one study mentions a web-archived news corpus of ~34 million words from a prominent newspaper [11]. More modern archives are easier: many Bangladeshi and Indian Bengali newspapers have online editions from the late 1990s or 2000s onward. **Prothom Alo**, for instance, is a top Bangladeshi newspaper with an online archive starting around 1999 – its articles over 20+ years would reflect ongoing language shifts. We can gather news text by web scraping (as done in some prior research [12] [13]). News articles provide structured, relatively formal language and cover a range of topics, which helps to observe words in varied contexts across time.

3. Other Written Sources: Beyond literature and news, we can include **magazines, journals, or official documents**. For example, Bangla Academy publications or government bulletins from the 1960s–1980s might contain useful text. If accessible, transcripts of radio/TV programs or subtitles could also reflect language of those times. The availability is hit-or-miss – we may need to rely on what libraries and online repositories have digitized. Collaborative projects like the **EMILLE corpus** (from early 2000s) included some Bengali text (about 2 million words from newspapers) [1], but that covers only a slice. More recently, the **IndicCorp** project compiled 3.9 million Bengali news articles (~836 million tokens) from the web [2], and the **OSCAR** corpus (Common Crawl data) contains ~632 million Bengali words from assorted internet text [2]. These large corpora skew towards the 2010s (since web content is mostly recent), but they can be treated as representing "modern" Bengali usage. The **KUMono corpus** (2022) is another comprehensive Bengali corpus (350+ million words) built by scraping 18 categories of websites [14] [15] – it presumably includes news, blogs, literature, etc., though mostly contemporary online text.

4. Social Media and Informal Text: To capture unstructured, colloquial usage, one should look at forums, social networks, and other user-generated content. Bengali is used on platforms like **Facebook, YouTube comments, Twitter**, and regional forums. These can reveal recent slang or semantic shifts that might not appear in formal writing. For instance, Bangla internet slang often incorporates new meanings for words or borrowings from English. A challenge is that Bangla social media text may be written in the Bengali script or romanized (transliterated into Latin script), and it's usually very noisy (inconsistent spelling, code-mixing with English, etc.). Nevertheless, including a **social media corpus (2010s–2020s)** would help the model detect emerging usages. Some researchers have compiled such data for tasks like sentiment or hate speech detection – for example, one survey cites Bengali hate speech datasets from Facebook comments and YouTube, which could be repurposed as a corpus [16]. Also, the **BanglaBERT** project collected a huge 27.5GB corpus from websites, including blogs and social media text [17], which indicates there is ample raw data if one can gather it. For our project, even a few million words of recent informal text would complement the more formal corpora.

In summary, **corpus feasibility** looks positive. We would combine: (a) *historical text* (literature and any archived news we can get for 1950s–1990s), with (b) *modern text* (2000s–2020s news, web, and social content). If older digitized texts are limited, an alternative is to lean on literature for mid-century language and treat 2000s news as the "later" period – this still allows a diachronic comparison. Importantly, all corpora should be cleaned and segmented by time periods (e.g., 1950–1970, 1970–1990, 1990–2010, 2010–2025) to enable analysis of changes over those spans. We may need to normalize the script encoding (Unicode normalization) and standardize old vs. new spellings. Once we have these corpora, we can proceed to modeling semantic change.

Techniques for Detecting Semantic Change

The project will integrate **n-gram statistics** with more advanced **embedding-based models** to detect and quantify semantic shifts. Each technique has a role: n-grams provide a transparent view of usage trends, while embeddings capture deeper contextual meaning. We outline the approaches and how they fit together:

N-gram Trend Analysis

Using n-grams involves looking at word occurrence frequencies and co-occurrences across time. For each target word, we can track its **frequency trajectory** over the decades and also identify its common **collocates** (neighboring words) in each period. For example, suppose the Bengali word for "wireless" (now often meaning *mobile phone*) was rare before 1990 but surges thereafter – a frequency spike might indicate a new concept or meaning entering usage. We could plot frequency vs. year (similar to Google Books Ngram Viewer graphs) for an initial sense of change. Moreover, changes in *collocations* can signal semantic shift: if the word "কুল" (phonetic *kul*) historically appears mostly in contexts about *lineage or family (as in kulin, kul-parampara)* but in recent years is frequently found with words like "ফ্যাশন" (fashion) or "স্টাইল" (style), it suggests "কুল" is now also used as a transliteration of English "cool" (meaning *trendy*) – a new sense. Such collocational shifts are detectable by comparing **top n-gram contexts** in the old vs. new corpora.

We plan to build or utilize an **n-gram viewer** for Bengali (analogous to Google's, but on our corpora). In fact, researchers in Bangladesh have already created a prototype called *Pipilika N-gram Viewer*, trained on a large Bengali newspaper corpus [18]. That tool allows querying a word and seeing its frequency over time. We can adopt a similar approach: index our time-stamped corpus and pre-compute n-gram counts. For each candidate word, we will examine:
- *Relative frequency change:* Did the word become significantly more or less common? (This can hint at semantic change if, say, a new meaning made a word suddenly popular.)
- *Collocate shift:* What other words appear most with it in early vs. later periods? For instance, an adjective that shifts domain (e.g., from describing weather to describing people's attitudes) will show different neighbors in text.

This n-gram analysis serves two purposes: **(a)** to generate hypotheses (we might flag words with big context changes as potential semantic shifters), and **(b)** to provide interpretable evidence supporting the more complex models. However, n-grams alone have limitations – frequency changes might reflect topics or censorship rather than meaning, and they won't directly tell us if the sense changed (a word could maintain meaning even if frequency drops). Thus, we integrate them with embedding-based methods for a fuller picture.

Static Word Embeddings and Alignment

Static word embeddings like Word2Vec or fastText represent each word as a fixed vector capturing its general context usage. The plan is to train separate embedding models for each time period's sub-corpus, then **align** these models to compare the vectors of the same word across time. This approach is well-established in semantic change detection [19]. Essentially, we create, say, a 1950s embedding space and a 2020s embedding space; if a word's vector "moves" far between these spaces, it implies its contexts (and likely meaning) have shifted.

A key technical step is **vector space alignment**: embeddings from different corpora are not directly comparable because the coordinate axes are arbitrary after training. We solve this by using an alignment algorithm (typically an Orthogonal Procrustes transformation) to rotate/translate one space to best match the other [20]. Hamilton et al. (2016) did this by learning an orthogonal matrix that maps the embedding from one year to another while preserving relative distances [20]. After alignment, we can compute similarity metrics – for each word, the **cosine similarity** between its old vector and new vector indicates how stable its meaning stayed. A low cosine similarity (or high distance) means the word's contextual meaning changed a lot. We can also look at **nearest neighbors**: for example, if in 1970 the nearest neighbors of "মাউস" (*mouse*) are "ইঁদুর" (*rat*, similar animal) but in 2020 they are "কম্পিউটার" (*computer*) and "কর্কর" (*click*, etc.), it shows a semantic shift from animal to computer accessory. Tools like these give a quantitative change score per word and even hint at the new vs. old meanings via neighbor lists.

Several embedding algorithms could be tried: **Word2Vec (Skip-gram or CBOW)**, **GloVe**, or **fastText** (which handles morphology by subword vectors). fastText might be advantageous for Bengali because it can form representations for rare inflected forms using character n-grams, alleviating sparsity. We will likely start with Word2Vec/fastText on each time-slice corpus (ensuring each has sufficient size – we might combine decades if needed to get robust training data). After training and aligning, we will rank words by how much they moved in vector space. Many past works define a threshold on cosine distance or a statistical test to decide if a word "underwent significant change" [19]. We can do the same for evaluation purposes. Importantly, this method can detect *semantic drift* even for words that did not dramatically change frequency. It captures subtle shifts: e.g., the word "ডাক্তার" (*doctor*) might always be frequent, but its context in 1950 might emphasize "গ্রাম" *(village)* and "হাকিম" *(practitioner)*, whereas in 2020 it appears with "সার্জারি" *(surgery)* and "চিকিৎসা বিজ্ঞান" *(medical science)*, reflecting professionalization of the role – a semantic broadening.

One limitation to note is that static embeddings conflate all senses of a word into one vector. If a word gained a new sense but still retains the old one, the vector shift might be modest (since the embedding averages all contexts). To mitigate this, researchers sometimes use **post-hoc sense clustering**: e.g., cluster the word's contextual instances into two groups corresponding to senses. But a simpler strategy within static embeddings is to use **local neighborhood** measures: compare not just the word's self-similarity, but the overlap in its top-k neighbors between times [21]. A large change in neighbors (even if the word remains somewhat similar to itself) can reveal semantic change. We will incorporate such metrics (like neighbor intersection size or rank shifts) as robust indicators.

Contextual Embeddings and Language Models

In recent years, **contextual embeddings** (from models like BERT, ELMo, or GPT) have opened new ways to detect semantic change. These models generate a *different* vector for a word depending on the sentence context, effectively capturing word sense distinctions. For a semantic change task, one approach is to take a pretrained language model and feed it sentences from different eras, then analyze how the word's contextual representations differ by era. For example, we could take a Bengali BERT model (such as **BanglaBERT** [17] or multilingual BERT) and input all instances of a target word from the 1960s corpus and from the 2020s corpus. By averaging those vectors or clustering them, we can see if the 2020s usages form a distinct cluster separate from the 1960s usages – which would indicate a meaning change. Some studies cluster contextual embeddings of target words and then compare cluster centroids across time to detect new senses [8].

Another strategy is to fine-tune a language model on each time period's data (essentially creating a "1960s BERT" and a "2020s BERT") and then use those models to embed words or predict word meanings. A Romanian study trained ELMo (a contextual embedding model) separately on old vs. new corpora and found it effective in capturing meaning differences [6]. Fine-tuning BERT for each period might be computationally heavy and possibly overkill for a thesis; instead, using one model and extracting context embeddings may suffice. One concrete method: compute a **contextual distance** measure – for instance, take the set of BERT embeddings for all instances of word *W* in period A and period B, and calculate how far apart these sets are (e.g., using average cosine similarity or a Wasserstein distance between the two distributions). If the distance is large, the word's usage has changed. This can complement static embeddings by handling polysemy: if a new sense appears, half the contexts might cluster separately.

Contextual models can also be leveraged to identify *which sense* changed. By examining attention weights or example sentences that are most indicative of a new sense, we could present qualitative insights. For example, if the word *"আইকন"* (*icon*) originally only meant a religious image and later also means a celebrity or computer icon, the model might show two distinct usage clusters – one around religious words and one around tech or pop culture words.

Integrating N-grams and Embeddings: The project can combine these approaches in a pipeline. First, n-gram and frequency analysis can highlight candidate words and provide intuitive visualizations (e.g., a graph of usage frequency or a table of top collocates per era). Then, for each candidate, static embedding shift can quantify the change, and contextual embedding analysis can validate it and perhaps distinguish multiple senses. The n-gram trends may also be used as features in a simple classifier: for instance, a word whose frequency and neighbor distribution changed significantly might be classified as "changed" by a supervised model (if we had a small labeled set of known changes vs. stable words for training). However, given likely lack of labeled data in Bengali, we will rely mainly on unsupervised detection and then manually verify with dictionaries or expert judgment whether the changes make sense.

Case Studies: Demonstrating Semantic Evolution

A compelling part of the thesis will be **case studies of specific Bengali words** that underwent notable semantic shifts. We will select a handful of target words and tell their "story" over time, backed by data. These case studies serve as qualitative validation of the modeling approach. Some possible examples:

- **Words that Gained New Meanings:** For instance, *"ডিজিটাল"* (*digital*) in Bengali was barely used in 1950 (if at all, perhaps only in technical contexts meaning *numerical*), but by 2020 "ডিজিটাল" is a buzzword meaning *modern, electronic,* or even *online* (as in "ডিজিটাল বাংলাদেশ" meaning a technologically advanced Bangladesh). The word's collocates might have shifted from mathematical terms to governance and lifestyle terms. Another example is *"ভাইরাস"* (*virus*). Historically it meant biological virus, but recently it's also used for computer viruses; additionally in the COVID era it took on a much more ubiquitous presence in everyday discourse. We can show how its frequency spiked in 2020 and how its contextual association expanded from medical journals to general news.

- **Semantic Pejoration or Amelioration:** Words sometimes change in connotation. A Bengali word like *"চালাক"* (which traditionally means *clever*) might have shifted in nuance – for example, older texts use it positively (*wise*), but in modern usage *"চালাক"* can imply *sly or crafty* (a negative tone). By

analyzing context, we could see a change in sentiment of surrounding words, illustrating a semantic drift in connotation rather than dictionary definition.

- **Loanwords and Global Influence:** Many foreign words have been absorbed into Bengali and sometimes change meaning in the process. The English *"viral"* is now used in Bengali (often written as "ভাইরাল") to mean *trending/popular*, beyond its original medical meaning – a change that occurred in the last decade. Similarly, "গুরু" in Sanskrit meant *teacher* (literally *heavy*), in Bengali it kept the *teacher* sense, but through English influence "guru" can now also mean an expert in a field (e.g., "ফিটনেস গুরু" for *fitness guru*). We can trace how often "গুরু" appears in contexts of religion/teaching versus skill/expertise over time.

- **Cultural and Political Shifts:** The word "স্বাধীনতা" (*swādhīnatā*, meaning *independence/freedom*) before 1971 was a general concept, but after the Bangladesh Liberation War, "স্বাধীনতা" took on the specific historical meaning of the independence of Bangladesh (often with a capital "S" implied). Its usage after 1971 is often in commemorative contexts, and phrases like "স্বাধীনতা আন্দোলন" (freedom movement) became common. While the core meaning didn't change, the **contextual focus** did – our models might detect that new associations (like names of war heroes or dates) appear post-1971, effectively a shift in the word's semantic network.

For each such case, we will present evidence: e.g., a small table of example sentences from different decades, the word's nearest neighbors in embeddings from old vs. new corpora, frequency graphs, etc. If possible, we might include visualization (like a timeline or a clustering plot) to make the semantic trajectory clear. These case studies will illustrate the real-world validity of the computational findings and also be an interesting narrative in the thesis. They also help in evaluation: if our methods flag these words and correctly characterize their change, it boosts confidence in the approach. Conversely, if a known changing word is missed by the model, it prompts us to refine our methods or data.

Challenges and Limitations

Working on semantic change in Bengali comes with a set of challenges, which we need to be mindful of and plan for mitigation:

- **Data Sparsity in Early Periods:** The farther back we go (1950s, 1960s), the less digital text is available. Our 1950s–1970s corpora might be relatively small in size compared to the 2000s corpora. This imbalance can bias models – e.g., a Word2Vec model trained on only, say, 2 million words from the 1950s will be less stable and lower-quality than one trained on 100 million words from the 2010s. Sparsity can lead to unreliable embeddings (high variance in vector due to low context examples). **Mitigation:** We can aggregate decades (e.g., treat 1950–1970 as one period) to increase volume. We can also use algorithms that are robust to smaller data, like PPMI matrices or fastText with subword information to handle rare words. Another idea is to initialize the older corpus embeddings with the vectors from the newer corpus (assuming the core meanings are similar) and then adjust with the older data – this might stabilize training. If extremely needed, we might include texts from the 1930s-40s (pre-1950) just to bulk up "old" corpus size, on the assumption that linguistic drift from 1930 to 1950 is relatively small in comparison to 1950 to 2020 changes.

- **OCR and Text Quality:** If we rely on OCR for scanned newspapers or books, the text may contain errors (misrecognized characters, etc.), especially given the complexity of the Bengali script and degradation of old print. Such noise can affect both n-gram counts and embedding training (e.g., "সরকার" becoming "সডকার" due to a smudge can create what looks like a new token). **Mitigation:** We will apply cleaning steps: use language models or dictionaries to correct obvious OCR mistakes, and possibly filter out low-frequency aberrations. Using modern font text (like Wikisource transcriptions) whenever possible is preferable to raw OCR. For embeddings, extremely low-frequency tokens anyway have little effect on the overall vector space – we can impose a frequency cutoff to ignore hapax legomena (words appearing only once).

- **Differences in Genre/Style:** When combining different sources, we must consider that changes we observe might stem from genre or register differences rather than true semantic change. For example, comparing literature from 1960 with social media from 2020 might reveal huge differences in word usage – but some of that is because spoken-style informal language is more present in social media. Words like "বাবা" (father) might appear formally in literature but in social media, "আব্বু" or "ড্যাড" might be more common – that's a lexical preference change, not a meaning change. **Mitigation:** We should ideally compare like with like. That means, if possible, also include *contemporary literature* to compare with classic literature, and *earlier informal language* (if available, e.g., transcripts of plays or interviews) to compare with social media. In practice, to keep it simple, we might focus the core analysis on a single genre (e.g., journalism/news language across time, or literature across time) to get a clean signal of semantic change. Then separately, we could analyze informal vs formal to see register effects. We will clearly document the composition of each period's corpus so readers know what differences might be confounding the results. Another mitigation is to use **embedding alignment on a genre-by-genre basis** – e.g., align old vs new literature embeddings and old vs new news embeddings separately. If a semantic change is real, it should show up in both comparisons.

- **Lack of Gold Standard Data:** Evaluating semantic change detection is tricky without a ground truth. For English, there are datasets where human annotators or dictionaries labeled which words changed meaning between two periods (for example, the SemEval "changed" vs "unchanged" word lists). For Bengali, we do not have an established list of words that have changed meaning over the last 70 years. This raises the question: how will we know if our model's outputs are correct? **Mitigation:** We will rely on indirect evaluation. The case studies with expert judgement (or consulting Bengali dictionaries from different eras, if available) can validate some changes. We can also do a small *survey or expert annotation* ourselves – for a random sample of words the model flags as changed, we could present example usages from early and late periods to Bengali speakers (or linguists) to confirm if they perceive a meaning difference. Additionally, known examples (like technological words or sociopolitical terms) serve as a partial gold standard; we expect those to appear high in our "changed" ranking. If they do, that's a positive sign. If our model is picking up mostly noise (like proper names or morphological variants), we would need to adjust it (for example, ensure we exclude named entities and focus on common nouns, verbs, adjectives which are more likely to undergo semantic shifts).

- **Polysemy and Sense Isolation:** Bengali words often have multiple senses (e.g., "কলি" can mean a flower bud, a decadent age (Kali Yuga), or a nickname). A word might add a new sense while keeping the old, making detection harder. Our static embedding approach might yield a moderate change score that's hard to interpret. **Mitigation:** Using contextual embeddings as mentioned can help

separate senses. We might also deliberately focus analysis on words that shifted **dominant sense** (i.e., the primary usage changed). If a word only gained a minor new sense, it might not be "important" enough to highlight in a thesis of this scope. As a computational solution, there are specialized methods like training **Gaussian embeddings or mixture models** that model a word as having multiple sense clusters and track those over time [22], but implementing those might be beyond the thesis scope. Instead, we will carefully interpret results and possibly do manual sense clustering on interesting cases.

- **Computational Resources:** Training multiple embedding models and especially any deep contextual models can be time-consuming. However, given the corpora sizes we anticipate (tens or a few hundred million words per period at most), Word2Vec training is quite feasible on a modern laptop or university server. Fine-tuning BERT on a large corpus is heavier; if needed we could use a subset or fewer epochs. We should leverage any pre-trained Bengali embeddings available (like fastText's Bengali vectors or BanglaBERT model) to save time – for example, we could start from those and fine-tune on specific periods to get "historical BERT" approximations.

Despite these challenges, none are insurmountable. The key is to acknowledge and account for them in the methodology. For instance, when presenting results, we will not only list words that the model found to have changed, but also provide the evidence (contexts, etc.) and note any potential confounds (e.g., "this word shows change, but it coincides with a genre shift in the data"). By being transparent, we ensure the conclusions are well-supported.

Novelty and Publication Potential

This thesis project appears to have strong **novelty** in the context of Bengali NLP. Lexical semantic change detection has been a trending topic in computational linguistics, but mostly for high-resource languages. Doing it for Bengali (and potentially other low-resource languages by extension) would be relatively new. A literature search yields virtually no prior comprehensive studies on diachronic semantic change specifically in Bengali, which means our work could be the first to present findings on this topic. That novelty alone – applying and evaluating these methods on Bengali – provides **publication potential**, especially if we also contribute resources (like a benchmark dataset or corpus) or insights unique to Bengali's linguistic evolution.

Here are a few aspects that can make it publishable in a computational linguistics venue:

- **New Data Resource:** If we manage to compile a sizeable diachronic corpus (especially if it includes older texts that are hard to find), that corpus itself is valuable. We could release a **Bengali Diachronic Corpus** with standardized splits (e.g., 1960s vs 2000s) and perhaps even some human annotations of semantic change for evaluation. Conferences like LREC (Language Resources and Evaluation Conference) or workshops on low-resource languages would appreciate this resource contribution. For instance, a paper could be framed as "First Bengali Semantic Change Corpus and Experiments".

- **Adaptation of Methods:** We will be applying known techniques (alignment-based static embeddings, contextual embeddings, etc.), but we might have to adapt them for Bengali's characteristics. Any novel twist – such as incorporating n-gram collocate information into the change scoring algorithm, or handling script/orthography changes – could be a methodological

contribution. Even a thorough comparison of static vs. contextual approaches on Bengali (similar to what was done for Romanian [6]) would be informative to the community, since it tests whether findings in English (like "contextual models outperform static ones") hold true for a very different language. If, for example, we discover that for Bengali, subword-aware static embeddings (fastText) detect certain changes better due to rich morphology, that's an interesting finding to report. This could be framed in a paper about "Semantic Drift in a Morphologically-Rich Low Resource Language".

- **Linguistic Insights:** The publication potential is higher if we not only apply models but also extract some *linguistic interpretations*. For instance, perhaps our analysis reveals that *social/political vocabulary* in Bengali changed more rapidly in 1970–1990 (due to the turbulent history around 1971) whereas *core vocabulary* remained stable – essentially testing Hamilton's observed laws in a new setting. If we confirm or contrast the **frequency law** or **polysemy law** in Bengali (e.g., do common words indeed change less often in our data? Do words with multiple meanings show more drift?), that directly ties into ongoing debates in semantic change research [23] [24]. We can cite those laws and discuss whether Bengali follows them or if there are cultural factors that create exceptions. For example, perhaps certain very frequent words did change meaning because of deliberate language reforms (like in Bengali, some Sanskrit-origin words were promoted over Persian-origin words in the mid-20th century, changing the usage of common terms). Such insights would interest venues focused on computational sociolinguistics or historical linguistics.

- **Low-Resource Focus:** There is growing interest in NLP for low-resource languages and how well methods generalize. A study demonstrating semantic change detection for Bengali could be submitted to conferences like ACL or EMNLP (main NLP conferences), especially if we emphasize the challenges of low-resource setting and how we overcame them. We might also compare Bengali with another language (maybe we could do a mini-study on Hindi or another Indian language if data permits, to show cross-language differences). That broader comparison could increase impact. But even focused on Bengali, it has merit as a case study from South Asia, a region underrepresented in such research.

To gauge publication potential, consider that a very recent paper was on Romanian semantic change [6] – a language with ~24 million speakers – published on arXiv and likely submitted to a conference, underlining that this topic is active. Bengali, with its 250+ million speakers, has equal if not greater need for such research. As long as our work is executed rigorously (with evaluation and clear findings), there's a good chance it could be published in a workshop or even a main conference session. We should aim to write up our results targeting venues like *ACL Workshops on Lexical Semantics or NLP for Similar Languages*, or journals like *Computational Linguistics (special issue on language change)* or *Language Resources and Evaluation*. Emphasizing the novelty, we will note in papers that "to the best of our knowledge, this is the first systematic computational study of lexical semantic change in Bangla." This signals a contribution of new knowledge.

One more angle: If our thesis yields interesting examples and perhaps identifies sociolinguistic trends (like influence of English on Bengali semantics, or differences between Bangladeshi and Indian Bengali usage over time), those findings could also be published in more linguistic-oriented venues or journals on South Asian languages. Interdisciplinary appeal (combining NLP and linguistic analysis) often strengthens a publication's value.

Conclusion and Next Steps

In conclusion, a thesis project on detecting and modeling semantic change in Bengali from the 1950s to present is **feasible and promising**. We have identified data sources that, when combined, will provide a reasonable diachronic corpus for analysis. We will leverage proven NLP techniques – n-gram analyses for transparency and word embeddings (static and contextual) for deeper insights – adapting them to Bengali's needs. While challenges like data sparsity and lack of benchmarks exist, we have outlined strategies to mitigate these, such as aggregating corpora, careful alignment, and qualitative case studies. The novelty of this work in the Bengali context means there is high potential for academic contribution. If successful, this research will not only produce an interesting thesis but could also lead to a publishable paper, contributing both a resource (Bangla diachronic data) and findings to the global research on lexical semantic change. Ultimately, beyond publication, the project will enrich our understanding of how Bengali words like *"cool"* (or its local equivalents) evolved from our grandparents' generation to today's – a fascinating journey of language through time.

References and Key Resources: (selection)

- Hamilton, W. et al. (2016). *Diachronic Word Embeddings Reveal Statistical Laws of Semantic Change*. ACL. [3] [20]
- Senapati, A. (2023). *A Semi-automated Approach for Bengali Neologism*. SN Computer Science. [4]
- Truică, C-O. et al. (2023). *Semantic Change Detection for the Romanian Language*. arXiv preprint. [6] [8]
- Kutuzov, A. et al. (2018). *Diachronic word embeddings and semantic shifts: a survey*. (Survey of methods)
- **Bengali Corpora:** Chaudhury et al. (2023). *VACASPATI: A Diverse Corpus of Bangla Literature*. arXiv. [10]; Akther et al. (2022). *KUMono 350M Word Bangla Corpus*. IEEE Access. [14]
- Relevant Tools: Pipilika Bengali N-gram Viewer (ICBSLP 2018) [18], BanglaBERT (EMNLP 2022) [17], fastText Bengali vectors (Grave et al. 2018).
- SemEval-2020 Task 1: *Unsupervised Lexical Semantic Change Detection* (for task design and evaluation ideas).

---

[1] [2] [10] [17] arxiv.org
https://arxiv.org/pdf/2307.05083

[3] [20] [23] [24] Diachronic Word Embeddings Reveal Statistical Laws of Semantic Change
https://aclanthology.org/P16-1141.pdf

[4] [12] [13] A Semi-automated Approach for Bengali Neologism | SN Computer Science
https://link.springer.com/article/10.1007/s42979-023-01866-2?error=cookies_not_supported&code=174ff9b7-e1b0-4e95-a567-41b5dc9c06cc

[5] (PDF) Semantic Change of Words Entered Into Another Language ...
https://www.academia.edu/105717427/Semantic_Change_of_Words_Entered_Into_Another_Language_Through_the_Process_of_Language_Borrowing_A_Case_Study_of_Arabic_Words_in_Benga
uc-sb-sw=5179744

[6] [7] [8] [9] [19] [21] [2308.12131] Semantic Change Detection for the Romanian Language
https://ar5iv.org/pdf/2308.12131

[11] [18] (PDF) BanglaLM: Bangla Corpus for Language Model Research
https://www.academia.edu/53113437/BanglaLM_Bangla_Corpus_for_Language_Model_Research

14 15 (PDF) Compilation, Analysis and Application of a Comprehensive Bangla Corpus KUMono
https://www.researchgate.net/publication/362502807_Compilation_analysis_and_application_of_a_comprehensive_Bangla_corpus_KUMono

16 Hate speech detection in the Bengali language - Journal of Big Data
https://journalofbigdata.springeropen.com/articles/10.1186/s40537-024-00956-z

22 A Word Sense Distribution-based approach for Semantic Change...
https://openreview.net/forum?id=oOKU31j9Q6