Lost in Transcription: Embeddings from ASR Transcripts

JONATHAN QUINN, Department of Computer Science, ETH Zurich, Switzerland

Supervisors: Prof. Ryan Cotterell, Dr. Alexander Hoyle

The proliferation of spoken-word corpora presents a significant opportunity for Natural Language Processing, yet the reliability of these datasets is compromised by artifacts from Automatic Speech Recognition (ASR). This thesis demonstrates that ASR errors extend beyond simple inaccuracies to include a severe form of data corruption, termed "looping hallucinations," where models invent and repetitively insert text with no basis in the source audio.

This work introduces a novel, scalable methodology to detect and mitigate these hallucinations in a large-scale podcast corpus. Following this data-centric sanitization, a comparative analysis of word embeddings trained on podcast transcripts versus written news reveals a fundamental divergence: the two media construct distinct "semantic realities" from the same historical events. Furthermore, an analysis of the podcast embeddings exposes a complex landscape of gender bias, where traditional stereotypes persist alongside novel associations that reflect evolving social roles.

Ultimately, this research argues that ASR-transcribed data cannot be treated as a direct substitute for written text. It provides a methodological framework for addressing ASR-specific artifacts and underscores the critical importance of data-centric inquiry, revealing that the "noise" in transcribed data can be as illuminating as the signal itself.

1 INTRODUCTION

1.1 Motivation

In the last decade, natural language processing has been transformed by word embeddings: dense vector representations of words that capture complex patterns of meaning directly from text. Foundational models like Word2Vec and GloVe, though now joined by larger contextual models, remain vital tools in computational social science for their simplicity, interpretability, and efficiency [7]. They provide one of the clearest windows into the subtle biases and social associations encoded in our language.

While written corpora such as news articles and web text have been studied extensively, the vast and growing world of spoken-language data remains comparatively under-explored. Podcasts, in particular, have become a central medium of public discourse, representing a unique communicative form: spontaneous and conversational, yet often highly produced [32, 41]. This medium introduces layers of complexity not found in clean, edited text. The reliance on Automatic Speech Recognition (ASR) systems to convert audio to text means that the resulting corpora are riddled with transcription errors, artifacts, and a peculiar failure mode known as "hallucinations" [6, 26, 36].

This raises a critical question: how do the unique properties of ASR-transcribed spoken language affect the word embeddings we learn from them? Do these embeddings replicate the semantic structures we expect from written text, or do they reveal new and systematic distortions? Answering this is essential for responsibly applying NLP methods to the millions of hours of spoken content being generated every day [23].

1.2 Research Questions & Contributions

This thesis was initially conceived to explore gender bias in the novel medium of podcasts. However, early explorations revealed that the foundational challenge was not in the analysis, but in the data itself. The ASR-transcribed corpus was so uniquely "messy" that a significant research pivot was necessary to first understand and tame the data before any meaningful downstream analysis could be performed. This journey from a specific application to a foundational data-centric investigation shaped the final research questions and contributions of this work.

This thesis makes the following contributions by addressing four central questions:

(1) **What are the unique challenges of preprocessing ASR-transcribed corpora compared to traditional written text?** I develop and document a scalable, parallelized preprocessing pipeline tailored to the quirks of ASR data, providing a practical guide for researchers working with similar spoken-word datasets. I also identify and rectify previously undocumented artifacts in both the podcast and news corpora used in this study.

(2) **How do ASR looping hallucinations manifest in large-scale podcast data, and how can they be systematically detected and mitigated?** I present a novel, database-driven methodology for identifying and quantifying a specific, under-explored type of ASR error: looping hallucinations. This analysis illuminates the prevalence and severity of this data corruption issue in the SPoRC podcast corpus.

(3) **How do word embeddings trained on spoken podcast transcripts differ from those trained on written news articles?** I conduct a rigorous comparative analysis of Word2Vec and GloVe embeddings from podcast and news corpora, examining their linguistic properties (lexical diversity), geometric structure (isotropy, principal components), and performance on a suite of intrinsic downstream tasks.

(4) **How is gender bias encoded in the conversational language of podcasts?** I provide one of the first in-depth analyses of gender bias in a large-scale podcast corpus, using clustering and analogy-based methods to reveal how the medium both reinforces traditional stereotypes and reflects evolving social roles.

1.3 Structure

This thesis is structured to guide the reader from foundational concepts to novel empirical findings.

- **Chapter 2** provides the theoretical background, reviewing the key literature on word embeddings, evaluation methods, bias measurement, and the specific challenges of ASR-transcribed data.
- **Chapter 3** details the two primary datasets, *SPoRC* (podcasts) and *NELA-GT-2020* (news), and describes the custom preprocessing pipeline developed to handle them.
- **Chapter 4** presents an in-depth analysis of ASR looping hallucinations and the framework engineered to detect and filter them.
- **Chapter 5** outlines the training process for the Word2Vec and GloVe embedding models on a high-performance computing cluster.
- **Chapter 6** offers a comprehensive evaluation and comparison of the resulting embeddings, analyzing their geometry and downstream performance.
- **Chapter 7** focuses on the investigation of gender bias within the podcast embeddings.
- **Chapter 8** discusses the broader implications of the findings, limitations of the study, and promising avenues for future work, before Chapter 9 concludes.

2 BACKGROUND & RELATED WORK

Understanding how language models learn patterns, relationships, and biases from text begins with a simple yet powerful idea: *words reveal their meanings through the company they keep.* This distributional hypothesis has guided linguistics and computational methods for decades [31], forming the bedrock of modern natural language processing. To situate the contributions of this thesis, this chapter guides the reader through four key areas: the foundational techniques for learning word embeddings, the methods for evaluating their quality, the tools for diagnosing their social biases, and the unique challenges posed by working with ASR-transcribed spoken-language corpora like podcasts.

2.1 From Counting Words to Learning Representations: Word Embeddings

For decades, representing words computationally involved simple counts: creating vast, sparse matrices of term frequencies (TF-IDF) or co-occurrence statistics (Latent Semantic Analysis) [19, 57]. While useful, these methods were limited. The breakthrough came with predictive models that learned to create dense, low-dimensional vector representations, or embeddings, that encode rich semantic and syntactic relationships in a continuous geometric space. Two architectures, in particular, catalyzed this shift: Word2Vec [45] and GloVe [52].

2.1.1 Word2Vec: Learning from Local Context. Developed by Mikolov et al. (2013), Word2Vec learns embeddings by training a shallow neural network on a simple predictive task. It comes in two main flavors:

- *Continuous Bag of Words (CBOW)*: Predicts a target word (w_t) from its surrounding context words (e.g., the two words before and after it).
- *Skip-gram*: Does the reverse, predicting the surrounding context words from a single central word. It is generally more effective for capturing semantics, especially with smaller datasets.

The objective for both is to adjust the word vectors to maximize the probability of observed word-context pairs. For the skip-gram model, this is expressed as maximizing the log probability over the entire corpus D:

$$\underset{\theta}{\text{maximize}} \sum_{(w,c)\in D} \log P(c|w)$$

This probability is typically calculated using the softmax function, which compares the similarity of the target word vector v_w and a context word vector v_c to the similarities with all other words in the vocabulary V:

$$P(c|w) = \frac{\exp(v_c \cdot v_w)}{\sum_{c'\in V} \exp(v_{c'} \cdot v_w)}$$

Because calculating this over the entire vocabulary is computationally expensive, Word2Vec employs efficient training strategies like *negative sampling*, where the model learns to distinguish true context words from a small number of randomly sampled "negative" (incorrect) ones. The result is a vector space where words appearing in similar contexts are positioned closer to one another, capturing nuanced semantic relationships.

2.1.2 GloVe: Global Co-occurrence Statistics. While Word2Vec learns from local context windows, GloVe (Global Vectors), introduced by Pennington et al. (2014), takes a different approach by explicitly modeling the co-occurrence statistics of the entire corpus. It learns word vectors such that their dot product equals the logarithm of their co-occurrence probability. The objective function is a weighted least-squares regression:

$$J = \sum_{i,j=1}^{V} f(X_{ij}) \left(w_i^\top \tilde{w}_j + b_i + \tilde{b}_j - \log X_{ij} \right)^2$$

Here, X_{ij} is the number of times word j appears in the context of word i, and $f(X_{ij})$ is a weighting function that down-weights very common and very rare co-occurrences. This formulation allows GloVe to encode not just proximity but ratios of co-occurrence probabilities, which helps it learn fine-grained linear relationships like analogies (e.g., king − man + woman ≈ queen).

Though newer contextual models now exist, these static embeddings remain widely used in sociotechnical and corpus linguistics research for their computational efficiency, transparency, and interpretability.

2.2 Intrinsic Evaluation: How Do We Judge Embeddings?

Once trained, how can we determine if a set of word embeddings is "good"? The standard approach is a suite of *intrinsic evaluations* [58] that measure how well the embeddings capture human-like semantic and syntactic structures. Common tasks include:

- *Word Similarity:* Measuring the correlation (e.g., Spearman's ρ) between the cosine similarity of two word vectors and human-judged scores of their semantic relatedness on benchmark datasets such as WS-353 [25] or MEN [2].
- *Analogy:* Testing whether vector arithmetic can solve proportional analogies such as "king is to man as queen is to woman":

$$v_{\text{king}} - v_{\text{man}} + v_{\text{woman}} \approx v_{\text{queen}}.$$

- *Clustering:* Assessing whether words naturally group into their semantic categories (e.g., animals, tools). Performance is often measured with metrics such as the Adjusted Rand Index (ARI) [33] and Normalized Mutual Information (NMI) [61], which compare the algorithm's clusters to ground-truth labels.

While these benchmarks are invaluable, their results are known to be sensitive to choices in preprocessing and the domain of the training corpus [21]. For a more qualitative view, visualization techniques like *Principal Component Analysis* (PCA) and *t-SNE* [63] are used to project the high-dimensional vectors into 2D or 3D, revealing the global and local geometric structure of the embedding space.

2.3 Probing for Bias: When Embeddings Learn Too Much

A foundational discovery in modern NLP is that word embeddings don't just learn semantics; they also absorb and can even amplify the social biases present in their training data [15, 16]. This spurred a wave of research into both measuring and mitigating these harms.

The cornerstone of bias measurement is the *Word Embedding Association Test* (WEAT), an adaptation of the psychological Implicit Association Test. WEAT quantifies bias by comparing the association of two sets of *target* words (e.g., male vs. female names) with two sets of *attribute* words (e.g., career vs. family). The association score between a word w and attribute sets A and B is defined as the difference in mean cosine similarity:

$$s(w, A, B) = \text{mean}_{a \in A} \cos(w, a) - \text{mean}_{b \in B} \cos(w, b)$$

The WEAT effect size is the difference between these scores for the two target sets, with statistical significance determined via permutation tests. The *SC-WEAT* variant [15] adapts this for smaller or specialized corpora, making it suitable for this thesis. While powerful, these metrics are sensitive to the choice of seed words [4, 12] and risk oversimplifying complex social constructs like gender [35, 38].

Numerous debiasing algorithms have been proposed, from geometric projection methods like "hard debiasing" [13] and Iterative Nullspace Projection (INLP) [56] to the design of explicitly gender-neutral embeddings [34, 65]. However, their effectiveness is debated, with some work suggesting they may only hide biases rather than truly remove them [28, 44].

2.4 The New Frontier: Spoken-Text Corpora and ASR

Most NLP research has historically focused on clean, written text. The proliferation of digital audio media, especially podcasts, has created a rich but messy new source of language data. The inherently conversational, diverse, and global nature of podcasts has inspired the creation of new large-scale corpora like *SPoRC* [40] and "100,000 Podcasts" [18], enabling research into the linguistic and social dynamics of spoken communication [42, 49, 59, 62].

Central to this work are advances in *Automatic Speech Recognition* (ASR). Modern end-to-end systems like OpenAI's *Whisper* [54] have dramatically improved transcription accuracy across various languages and conditions [3, 37]. However, ASR systems are far from perfect. They are prone to systematic errors and unique *hallucinations*: transcribed text that has no anchor in the source audio [5, 6, 26, 27, 36]. A particularly disruptive and under-studied error is *looping hallucinations* [39], where the ASR model repeatedly outputs the same phrase, often overwriting genuine speech, a phenomenon linked to background noise, silence, or non-standard accents [6, 36]. Understanding and addressing these artifacts is a critical prerequisite for any downstream linguistic analysis of ASR-transcribed corpora.

3 DATASETS & PREPROCESSING

3.1 Datasets

The empirical foundation of this thesis rests on two large-scale media corpora. This section introduces these datasets, describes their scope, and outlines the features relevant to this work. The first corpus, SPoRC, consists of podcast transcripts, while the second, NELA-GT-2020, is a collection of news articles.

3.1.1 SPoRC. The Structured Podcast Research Corpus (SPoRC) is a large-scale dataset developed for the computational analysis of the podcast ecosystem, addressing prior limitations in data availability and the technical challenges of analyzing audio media [40]. The corpus contains over 1.1 million English-language podcast transcripts from 247,000 distinct shows, collected via public RSS feeds in May and June 2020.

SPoRC provides rich multimodal information, including:

- Transcripts for all 1.1 million episodes, generated using OpenAI's *Whisper* ASR system [54].
- Audio features and speaker turns for a 370,000-episode subset. Speaker diarization (segmenting audio into speaker turns) was performed on this subset using `pyannote` [14].
- Metadata for all episodes, including inferred speaker roles (hosts and guests), category, duration, and publication date.
- Speaker-turn-level data, including transcribed speech segments with timestamps, audio features, and inferred roles where applicable.

SPoRC was constructed using a highly-parallelized processing pipeline and incorporates models for extracting host and guest identities. The dataset is intended to facilitate new research directions into this medium, such as the study of online communities and temporal trends, and is available for non-commercial use.

For this thesis, the 370,000-episode subset was used, totaling over 54 gigabytes in its raw format. After preprocessing, the corpus size was reduced to approximately 12 gigabytes. The significant computational demands of this pipeline necessitated the use of a supercomputing cluster for data preparation and model training.

3.1.2 NELA-GT-2020. The NELA-GT-2020 dataset is an updated version of a 2019 corpus designed for studying misinformation in news articles [29]. It comprises nearly 1.8 million news articles collected from 519 mainstream and alternative news sources throughout 2020. All articles are in English[1] and originate from various countries.

Key features of the NELA-GT-2020 dataset include:

- *Ground Truth Labels*: Source-veracity labels from Media Bias/Fact Check (MBFC) are included, assigning each source a reliability score.
- *Embedded Tweets*: The dataset contains over 410,000 tweets embedded within the articles, along with their associated metadata.
- *Broader Scope*: The collection covers political, health, and general news, including major events such as the COVID-19 pandemic and the 2020 US presidential election.

The dataset is released in multiple formats, including JSON, which allowed for the direct application of the preprocessing pipeline developed for SPoRC. This uniform approach minimized potential confounders and enhanced the methodological robustness of the subsequent comparative analysis.

[1] See 3.2.3.

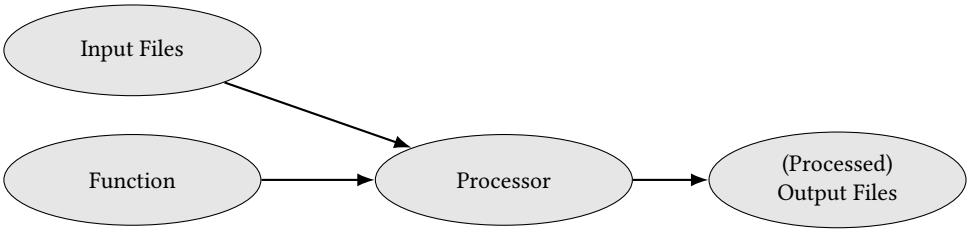

Fig. 1. First pipeline sketch.

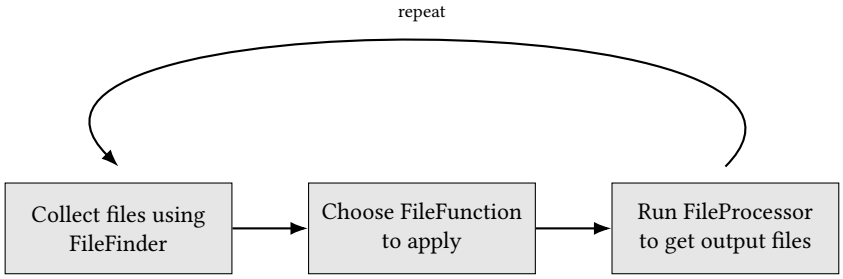

Fig. 2. Iterative preprocessing pipeline.

3.2 Preprocessing Pipeline

The preprocessing pipeline was designed primarily for the SPoRC dataset, the primary focus of this work, though its architecture was sufficiently general to be applied to the NELA-GT-2020 corpus as well.

3.2.1 Architectural Design. The decision to use the SPoRC dataset was motivated by its novelty and scale. Its considerable size (over 50 GB) presented a significant computational challenge. To align with the scaling hypothesis; i.e., that model performance improves with data volume; the entire corpus was processed, necessitating a scalable, parallelized pipeline.

During local development, memory constraints were managed by partitioning the corpus into smaller fragments. This motivated the design of a pipeline that applies a given function to a list of input files to produce a list of processed output files (figure 1). The core components of this architecture are

- `FileFunction`: An abstract class for operations that transform a single input file into an output file. Implementations include:
  - `EntrySimplifier`: Strips metadata from JSON entries.
  - `TextCleaner`: Removes unwanted text markers and artifacts.
  - `SentenceListCreator`: Converts raw text into tokenized sentence lists for Word2Vec.
  - `GloVeFormatter`: Prepares tokenized data for GloVe training.
- `FileProcessor`: A class that coordinates the application of a `FileFunction` across all input files.
- `FileFinder`: A class that collects a list of files from a directory based on specified criteria.

This modular design led to the iterative preprocessing workflow depicted in Figure 2. However, this architecture has limitations. It does not natively support one-to-many or many-to-one file operations, requiring that cross-file data dependencies be handled via caching. This created challenges, for instance, when merging podcast fragments that

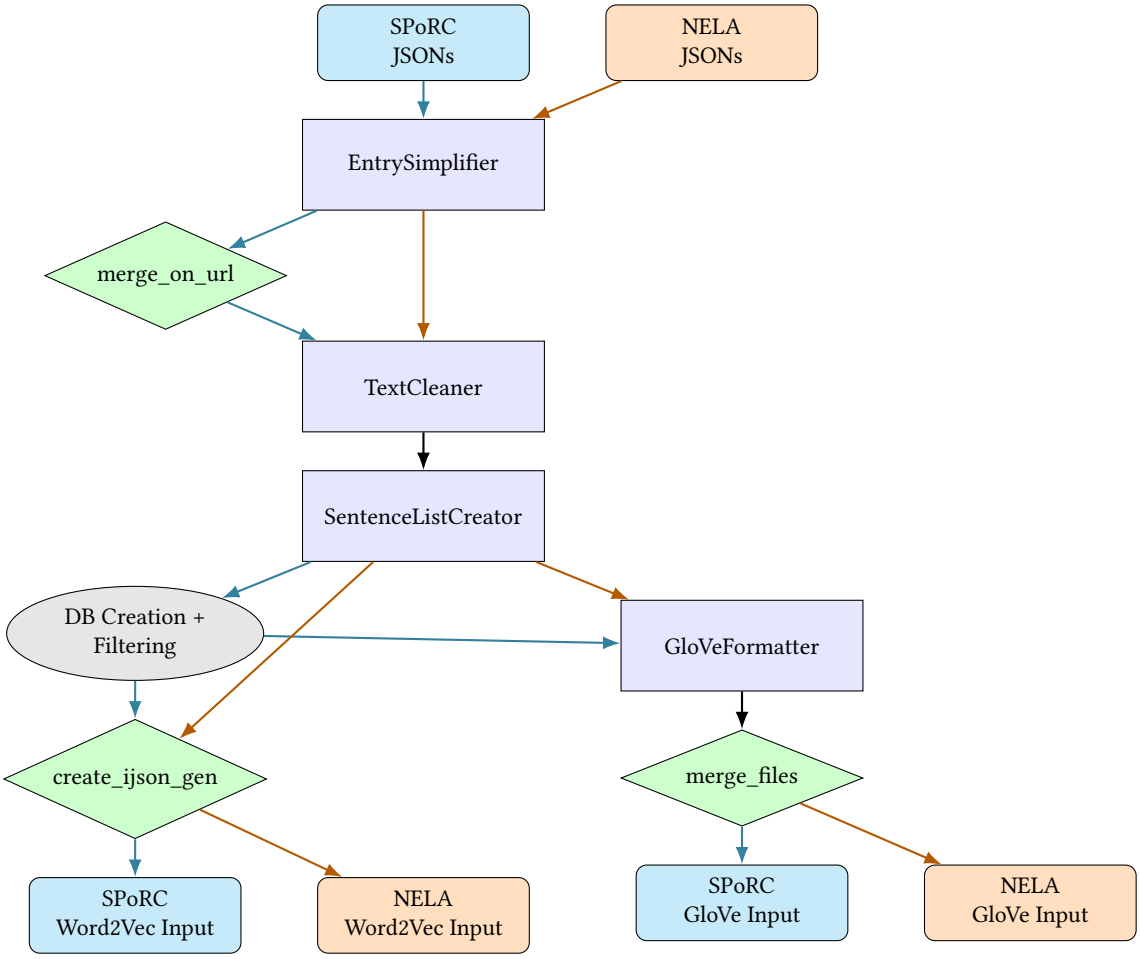

Fig. 3. Complete preprocessing pipeline.

spanned file boundaries, which required sequential processing and custom caching logic to ensure data integrity. These challenges are discussed further in section 3.2.3.

3.2.2 Final Preprocessing Pipeline. The finalized preprocessing pipeline is shown in Figure 3. The first stage used `EntrySimplifier` to remove all metadata irrelevant to dialogue tokenization, retaining only the URL and transcribed text for SPoRC and the main article text for NELA-GT-2020. An auxiliary method then merged SPoRC JSON fragments based on their source URL to reconstruct coherent, sequential podcast transcriptions. This step was critical because Word2Vec requires intact sentences as input.

The next stage, text cleaning, was the most iterative part of the process. The resulting `TextCleaner` class contains a long list of curated rules based on extensive corpus investigation. At this point, the data consisted of two sets of cleaned text files, with each line corresponding to a single podcast episode or news article.

To prepare the data for the embedding models, a two-step tokenization process was implemented for efficiency. As tokenization with `spaCy` was the primary computational bottleneck, it was run only once. The `SentenceListCreator` module processes each line of text, splitting it into chunks to accommodate `spaCy`'s internal memory limits, and then uses a `spaCy` pipeline to generate lists of tokenized sentences. A special stop token is appended to mark the end of each document.

From these intermediate JSON files, a sentence generator is created for Word2Vec training, and a separate plain-text version is created by the `GloVeFormatter` module for GloVe training.

Finally, the node labeled *DB Creation + Filtering* in Figure 3 represents a critical stage added after initial implementation. This step serves as an additional cleaning pass to address ASR-generated hallucinations in the SPoRC corpus. The details of this process are the focus of Chapter 4.

3.2.3 Pipeline Iterations & Lessons Learned. While the final architecture appears straightforward, its development involved significant iteration. The lessons learned may benefit future research on spoken-text data.

The primary architectural limitation was the `FileFinder–FileFunction–FileProcessor` pattern, which was not well-suited for one-to-many or many-to-one operations, such as the initial partitioning of the datasets or the final merging of files for GloVe. These steps required auxiliary methods with custom logic.

Text cleaning proved to be the most complex phase. The process of cleaning the SPoRC transcripts involved several key design decisions:

- *Non-speech audio*: Content within signifiers like parentheses or asterisks, used by Whisper to denote non-speech audio (e.g., music, laughter), was removed using `regex`.
- *Possessives and apostrophes*: To prevent inconsistent tokenization of named entities (e.g., "Obama" vs. "Obama's"), contractions were expanded using the contractions package, and possessives were removed. Remaining internal apostrophes (e.g., in "O'Reilly") were converted to underscores to align with the `spaCy` NER strategy.
- *Profanity markers*: `regex` rules were carefully designed to handle whitespace around asterisk-masked profanity (e.g., "f*ck") to avoid overly greedy pattern matching that could delete large, unrelated segments of text.

Initially, it was assumed that cleaning the written-text NELA-GT-2020 corpus would be simpler. However, a first pass with the podcast-optimized `TextCleaner` left a substantial amount of data contamination unaddressed within the NELA corpus. A qualitative analysis of the resulting embeddings revealed that several principal components were separating German-language tokens from English ones, indicating that the corpus was not exclusively English as documented.

This discovery prompted a manual, visual inspection of the NELA-GT-2020 data, which uncovered numerous web-scraping artifacts requiring removal:

- *Preambles and Suffixes*: Repetitive introductory and concluding text (e.g., slogans, copyright notices, social media links) were removed.
- *Web Scraping Dialogue*: Messages from news sites (e.g., subscription paywalls, JavaScript warnings) were stripped out.
- *Duplicate Articles*: A hash-based approach was used to identify and remove duplicate articles, a critical step for data hygiene.

The source of the foreign language content was traced to over 1,600 articles from a German news outlet, which were manually removed. This rigorous cleaning process highlights that web-scraped written corpora can contain as many, if not more, structural and content-related artifacts as ASR-transcribed spoken corpora [9, 21].

Finally, significant technical challenges were overcome in the implementation of `SentenceListCreator`. Performance issues related to `spaCy`'s memory usage and processing speed were addressed by moving from a naive implementation to a parallelized pipeline that processed text in smaller chunks. More subtle bugs related to the interaction between text cleaning, chunking, and stop-token handling required several rounds of debugging. The final, robust solution involved processing each document (line) individually and manually appending a stop token within the `SentenceListCreator`, thereby obviating the need for the legacy `StopTokenAppender` class.

3.3 Summary

In this section, I introduced the two large media corpora used for this thesis: SPoRC and NELA-GT-2020. I detailed the multi-step preprocessing pipeline designed to transform the raw, large-scale datasets into a clean and uniform format suitable for training embedding models. The core of this pipeline involved merging fragmented data, performing extensive text cleaning to remove artifacts from both speech-to-text transcription and web scraping, and tokenizing the corpora using a custom-built, parallelized architecture.

The process required significant iteration and optimization to handle the sheer volume of the data, as well as nuanced cleaning to manage irregularities like non-speech audio markers in podcasts and web-scraping noise in news articles. Ultimately, this meticulous preprocessing yielded two cleaner corpora, which are now ready to be used as input for the embedding models. The next chapter will discuss specific challenges uncovered during this process, namely the hallucinations present in the SPoRC corpus and the subsequent database creation and filtering steps implemented to address them.

4 HALLUCINATIONS

Likely transcription errors that report unspoken speech are referred to as “hallucinations," following recent literature that defines them as fluent outputs with little or no semantic connection to the source audio [27]. Variants such as “oscillation artifacts,” “repetitive hallucinations,” or “repetitive label errors” have also been used for specific subtypes, such as the repeated-sentence loops observed in Whisper ASR outputs [6, 26, 39].

4.1 Detecting Hallucinations

The analysis of nearly 169 million sentences across a 70-file split necessitated an approach that could handle large-scale data processing without overwhelming system memory. To address this, a PostgreSQL database was selected, leveraging its robust capabilities for efficient large-scale data management, complex queries, and indexing. The projected storage requirements for the database, in excess of 20 gigabytes, also prompted the use of a 2-terabyte external SSD to support preprocessing iterations for both corpora.

Given the scale of the SPoRC corpus, a direct storage and comparison of full sentences would be computationally and memory intensive. To overcome this, a more efficient method was implemented: each sentence string was fingerprinted using the xxHash algorithm. This approach, which creates a fixed-length hash for each sentence, conferred several key benefits:

- *Efficiency*: To find instances of looping, my future database operations would need to compare subsequent sentences to check if they match. Comparing fixed-length hashes is much faster and more memory-efficient than comparing variable-length strings.
- *Indexing*: Hashes can be easily indexed in the database, which dramatically speeds up the process of finding all occurrences of a specific sentence.
- *Data Integrity*: The fact that xxHash is quick made ensuring integrity much easier. Before actually removing a hallucinated sentence from the corpus, I recomputed the hash to ensure there were no mismatches.

The overall strategy was a multi-step process consisting of database population, analysis, and filtering. The next subsections go into more detail regarding database organization and hallucination detection methodology.[2]

4.1.1 Hash-based Database. The foundation of this analysis is a database table designed to index every sentence in the corpus. This initial step is handled by a script that populates a table named sentences with columns for sentence hash (`hash`), file number (`file_num`), line number (`line_num`), and sentence number (`sent_num`). This structure allows for the unique identification and location of any sentence across the entire dataset. To ensure data integrity, a unique constraint was added on `(file_num, line_num, sent_num)`. For performance optimization when dealing with millions of records, inserts were batched using the `psycopg.Connection`’s `executemany` method to reduce the number of individual database transactions.

4.1.2 Run-length Detection. Identifying repeated hashes was a necessary but insufficient step. To detect looping behavior, it was essential to find instances where identical sentence hashes occurred in long, consecutive runs. Manually querying each of the 5.5 million repeated hashes was infeasible.

To solve this, my `create_runs.py` script implements a two-step process:

(1) *Identify Frequent Hashes*: First, run a query to find all sentence hashes that appear more than a minimum number of times (given by `threshold` parameter) in the entire dataset. This reduces the search space considerably.

[2] As always, the referenced code can be found in my bachelor’s thesis repository [53].

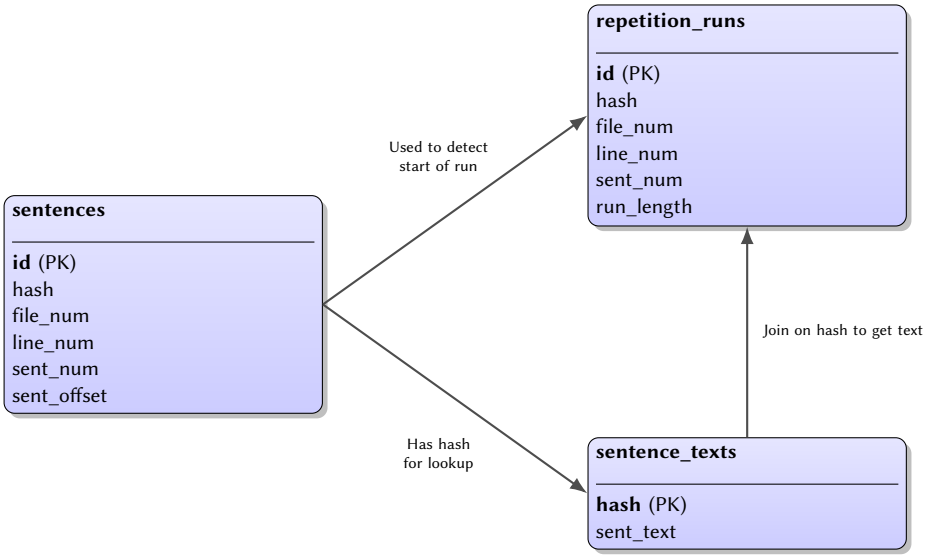

Fig. 4. Database schema.

(2) *Stream and Detect Runs*: For each of the frequent hashes, stream its occurrences, ordered by location (`file_num`, `line_num`, `sent_num`). By iterating through this ordered stream, the Python script can easily check if sentences are consecutive (e.g., if `sent_num` increases by one while `file_num` and `line_num` stay the same) and thereby measure the "run length" of this loop.

The results were stored in a new, much smaller table named `repetition_runs`.[3] A repeated run was only added if its run length was at least `threshold`. The new table stores only the starting location (and hash) of each detected run and its length. This approach is far more efficient than recalculating this information on the fly and made subsequent analyses faster and more repeatable.

4.2 Analysis

The database construction enabled the efficient location and removal of a specific type of hallucination: repeated sentence runs. While the schema and the SPoRC corpus's extensive metadata could enable more extensive analysis, the scope of this thesis is limited to classifying and quantifying this specific phenomenon. The following subsections detail the criteria used to classify sequences of repeated sentences as hallucinatory and document the extent to which the SPoRC corpus is affected.

4.2.1 Choosing a Threshold. To distinguish natural repetitions in human speech from machine-generated hallucinations, a clear threshold for run length was required. While preliminary hypotheses suggested a threshold of 3 or 4, an empirical approach was necessary to validate this assumption. During exploratory analysis, a key metric was identified: the uniqueness ratio, which tracks the proportion of unique repeated sentences relative to the total number of repetitions at a given run length. Plotting these ratios, as shown in Figure 5, reveals a crucial indicator for distinguishing between the two phenomena.

[3] At this point, I hadn't investigated hallucination-related literature all too thoroughly and was thus still referring to the looping phenomenon as "repeated sentence runs."

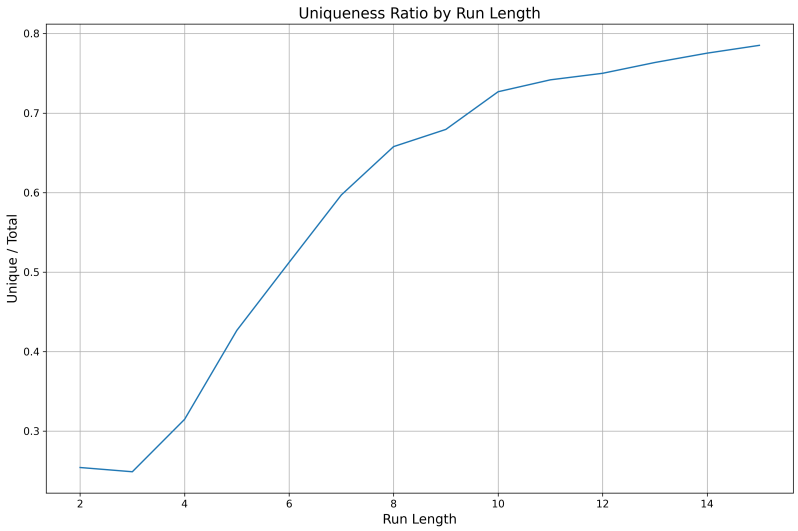

Fig. 5. Unique vs. total hash ratios per repeated run length.

To illustrate, consider repeated sentence runs of length 2. If a corpus contained two unique examples of such runs, such as *"Bear! Bear!"* and *"Ow! Ow!"*, and the former appeared only once while the latter appeared three times, the uniqueness ratio for run length 2 would be calculated as $2/(1+3) = 0.5$.

As you may notice, Figure 5 shows a distinct "elbow" at run length 3. For run lengths 2 and 3, the uniqueness ratio drops slightly, but stays around 0.25 for both. But after that, it begins to rise significantly, starting at around 0.32 for run length 4 and approaching 1 as run lengths increase. Why might this be the case?

In my view, the distinct elbow at a run length of 3 is a crucial indicator of the transition between natural repetitions in human speech and machine-generated hallucinations:

- *Run Lengths 2–3*: As illustrated, humans naturally repeat phrases for emphasis or as conversational filler. In podcast transcriptions, short repetitions like "Thank you, thank you" or "Yeah, yeah, yeah" are frequent and authentic features of spoken language. The low uniqueness ratio for these lengths reflects this; many different podcasts will contain these same common, short, repeated phrases.
- *Run Lengths 4+*: A repetition extending beyond three sentences is where the data likely transitions from human to machine origin. While a natural human utterance might, in rare cases of extreme emphasis, extend to four or five repetitions (e.g., "No, no, no, no!"), the probability of such an event decreases rapidly. The likelihood of a human speaker naturally repeating an entire sentence 200 times, for example, is effectively zero. Therefore, longer runs are almost certainly a signature of Whisper's "looping" behavior. Because these machine-generated sequences are less predictable than common human interjections, each long run is more likely to be unique, causing our ratio metric to rise toward 1.

The resulting hypothesis is that the elbow in Figure 5 represents the transition point from natural human repetition to machine-generated hallucinations. However, for the reader, a follow-up question may emerge: "If runs of length 4 and up are hallucinated and thus don't reflect common linguistic repetitions, why doesn't the ratio rise much quicker once we pass the threshold? Why doesn't it instantly shoot to one?"

A probabilistic explanation for this gradual rise is rooted in the frequency distribution of the loops themselves. The chance of a "collision," where two unrelated looping events happen to share the same sentence content, is a function of sample size. The data reveals tens of thousands of short runs (e.g., length 4), but only a handful of very long ones.

This distribution is likely a product of the ASR model's internal mechanics; the probability of terminating a repetitive loop may increase with each sentence, making extremely long runs rare events. Consequently, with a massive sample of short runs, it is statistically more probable that some common phrases will be generated independently in different contexts, lowering the uniqueness ratio. Conversely, as run length increases, the number of instances plummets, making a content collision between two distinct events extremely unlikely and pushing the uniqueness ratio toward 1.

4.2.2 Severity. This section documents the severity and extent of the looping phenomenon in the SPoRC corpus. The analysis of the 370,000-episode subset revealed a total of 104,376 repeated sentence runs of length 4 or more. The length distribution of these runs is presented in Figure 6.

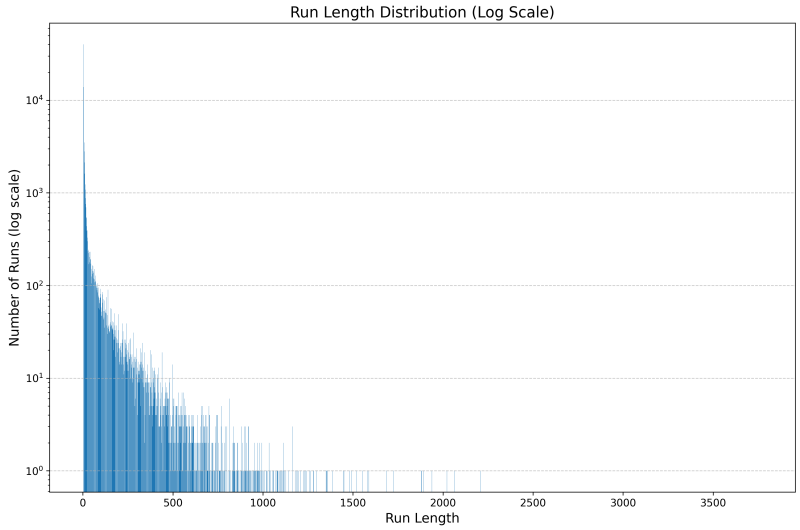

Fig. 6. Number of repeated runs per run length.

More specifically, looping hallucinations span from 40,119 runs of length 4 to a maximal length of 3,768 identical, repeated sentences in a single episode. This maximum is an extreme outlier; fewer than 13% of looping instances exceed 50 repetitions, and only 107 runs reach a length of 1,000 or more. The average run length is 32.76 sentences.

Although these hallucinations appear in only 7,516 distinct episodes (approximately 2% of the corpus), episodes with looping behavior are characterized by an average of 15.94 instances per episode. This suggests that the phenomenon is rare but intense within the affected episodes. Despite this intensity, the overall impact on the corpus is minor; a comparison of character counts reveals that all looping hallucinations collectively constitute only 0.959% of the corpus.

The most commonly looped sentences and their total repetitions are compiled in Table 1. As is evident from the table, repeated sentences tend to be short; the average sentence length in these runs is 22.38 characters (4.88 words). A noteworthy observation is the prevalence of self-prohibiting phrases, such as "I am not going to do that," which warrants further investigation into the ASR model's failure modes. The longest sentence found in a repeated run (of length 4) is 39 words long, demonstrating the diverse nature of these artifacts.

Sentence text	Reps
yeah	132161
i think that is a good thing	38320
i am not going to do that	37500
i am not going to say that	32869
i am not going to do it	19672
you	17975
and i think that is a good thing	15848
no	14067
i think that is a great thing	11873
i think it is a good thing	11554

Table 1. The most commonly looped sentences.

4.2.3 Remarks. In addition to the quantitative analysis, a qualitative examination of specific hallucination instances provided critical context. The existence of these looping hallucinations was first observed during a manual inspection of text files, where a long repetition of the sentence "I'm not going to get out of this." was identified. Subsequent investigation traced this instance back to its original audio source, revealing that these loops do not merely add content but actively overwrite existing audio segments. This finding suggests that looping hallucinations are not just an artifact of non-speech audio but can also represent a loss of data.

An initial hypothesis was that this phenomenon might be a form of censorship, as the overwritten passages often contained emotionally charged language or sensitive content. However, a more comprehensive review of the corpus suggests that looping is primarily an artifact of the ASR system's behavior when confronted with degraded or silent audio, a conclusion consistent with existing literature [36] and other research on ASR systems.

The discovery of this phenomenon also highlights the element of chance in research. This particular hallucination was only observed because it was located at the start of the first text file, a result of an unexpected file-splicing error in the preprocessing pipeline. Precisely because these hallucinations comprise less than 1 percent of the total corpus by character count, this serendipitous discovery underscores the importance of qualitative data inspection in large-scale quantitative studies. While this investigation diverged from the initial research proposal, the engineering, analysis, and data provided here establish a reproducible methodology for detecting and removing a significant source of data contamination. This work provides a valuable resource for future researchers utilizing the SPoRC corpus or other Whisper-transcribed datasets.

4.3 Summary

In summary, this chapter detailed the development of a robust, database-driven methodology to detect and quantify repeated sentence hallucinations in the SPoRC corpus. By leveraging sentence hashing and analyzing a novel uniqueness ratio, a threshold of a run length of 4 or more was established as the criterion for classifying a sequence as a hallucination. The analysis revealed that while these looping errors are rare, affecting only approximately 2% of episodes, they are intense within those affected episodes. Despite this intensity, their overall impact on the corpus is minor, accounting for less than 1% of the total character count. Qualitatively, these hallucinations were found to actively overwrite existing content, representing a form of data corruption that warrants careful consideration in any ASR-transcribed corpus analysis.

5 EMBEDDING MODEL TRAINING

Training word embedding models on large corpora is computationally intensive. Word2Vec requires significant processing time for iterative training, while GloVe is memory-intensive due to its reliance on a global co-occurrence matrix. Given the scale of the SPoRC and NELA-GT-2020 corpora, these computational demands made training on a personal laptop prohibitive. Consequently, a high-performance computing (HPC) cluster was utilized for all model training and the more demanding preprocessing tasks, such as tokenization.

5.1 Tools & Training Procedure

Word2Vec models were trained using the `gensim` library [67], an open-source package designed for efficient, scalable topic modeling and vector space model implementation. Its `Word2Vec` class is a `cython`-optimized implementation [8] of the original algorithm presented by Mikolov et al. [45], and its support for streaming large corpora from disk was essential for this project.

For GloVe models, the original C implementation provided by Pennington et al. [60] was used.

The training workflow was managed using `slurm` job submission scripts. For each model, computational resources were allocated based on corpus size and model requirements (specifics are detailed in the Appendix, Table 12). To monitor convergence and diagnose potential issues during training, logging was implemented to output progress to `slurm` reports.

5.2 Hyperparameter Selection

To ensure comparability between models, several key hyperparameters were held constant. The vector dimensionality was set to 300, a common choice that balances semantic expressiveness with computational efficiency. To manage vocabulary size and exclude infrequent words, the minimum word count (`min_count` for Word2Vec, `VOCAB_MIN_COUNT` for GloVe) was set to 5 for all models.

For the Word2Vec models, the skip-gram architecture (`sg=1`) was selected over CBOW for its superior performance on semantic tasks. A window size of 5 and negative sampling with 5 "negative" examples were used, which are standard values for this architecture and corpora of this scale. A subsampling parameter of `1e-5` was chosen to downsample highly frequent words, and models were trained for 5 epochs to ensure convergence.

The GloVe models were trained with hyperparameters based on the authors' recommendations. A symmetric context window of 15 was used to construct the co-occurrence matrix; this broader window allows GloVe to capture more global co-occurrence statistics than Word2Vec. The `X_MAX` parameter was set to 100, and models were trained for 50 iterations, both standard values for corpora of this size.

It is important to note that the architectural differences between Word2Vec (local context window) and GloVe (global co-occurrence statistics) mean that their embeddings are not directly comparable from a methodological standpoint. Architectural choice must be considered a confounding variable in the subsequent analyses. However, this limitation does not invalidate the findings; rather, comparing how these different architectures represent the same corpora provides valuable insight into model-specific properties. The results in chapters 6 and 7 are interpreted within this framework.

6 EVALUATION

This chapter presents findings on the linguistic characteristics, geometric structure, and downstream performance of the learned embeddings across both corpora and model types.

6.1 Embedding Space Geometry

6.1.1 Lexical Diversity & Word Frequency Distributions. To evaluate the lexical diversity of each corpus, this analysis uses the type-token ratio (TTR), which is the ratio of unique words (types) to the total number of words (tokens). A higher TTR indicates greater lexical variety, while a lower TTR suggests more repetitive language. Table 2 presents the TTR and other corpus statistics. Note that the podcast numbers are averages across the 11-fold split of the corpus, and the metrics are representative for both Word2Vec and GloVe models due to a shared tokenization process and `min_count` hyperparameter.

Corpus	Unique words	Total words	TTR
Podcasts (avg)	117,895	203,363,275	0.0005797
NELA	193,739	159,709,390	0.0012131

Table 2. Type-token ratio (TTR) and corpus statistics.

As shown in Table 2, the NELA corpus is substantially more lexically diverse than the SPoRC corpus. Even though it has over 40 million less total tokens than the average podcast model, the NELA Word2Vec model contains around 194,000 unique tokens, almost 76,000 more than its speech-to-text counterpart. This staggering difference is captured in the TTR values: NELA's TTR of 12.1×10^{-4} far exceeds the average podcast model's 5.8×10^{-5} TTR value.

There are multiple potential explanations for this. The first, and most natural, is that written language tends to be more varied and precise than spoken language. Spoken language is often more repetitive, featuring recurring greetings, pragmatic markers, filler words, and mid-utterance reformulations. In general, people's active spoken vocabulary tends to be more limited than their written one [47]. In contrast, news articles cover a broad range of global events and use lexical diversity as a function of clarity and precision. Podcasts may span diverse topics, but the genre's communicative conventions (e.g. hosts regularly greeting guests) and the live nature of the recordings (which leads to improvisation and fillers) result in a more repetitive and constrained vocabulary. Taken together, these properties of spoken discourse provide a natural explanation for the much higher TTR in the NELA corpus [17, 30].

The podcast corpus's larger total word count, despite its lower lexical diversity, is likely a consequence of the same phenomenon: spoken discourse tends to use shorter, less information-dense words more frequently, whereas written text relies on longer and more precise terminology. The token length data backs this up: 11.72 characters in the NELA model versus a mean of 8.69 across podcast models.

Figure 7 compares word frequency distributions for the podcast and NELA corpora. The first shows a classic Zipfian [66] distribution, while in the second, I log-scaled both axes to enable a sharper analysis. Also note the difference in frequency ranges: the first plot shows the top 2-150 most frequent words, while the second shows ranks 2 to 2000.

The visualization begins at rank 2, excluding the most frequent token ("the"). This token's extreme frequency, particularly in the NELA corpus, skewed the y-axis to a degree that rendered the plots uninformative for other high-frequency words. Its omission allows for a more meaningful analysis of the distributions.

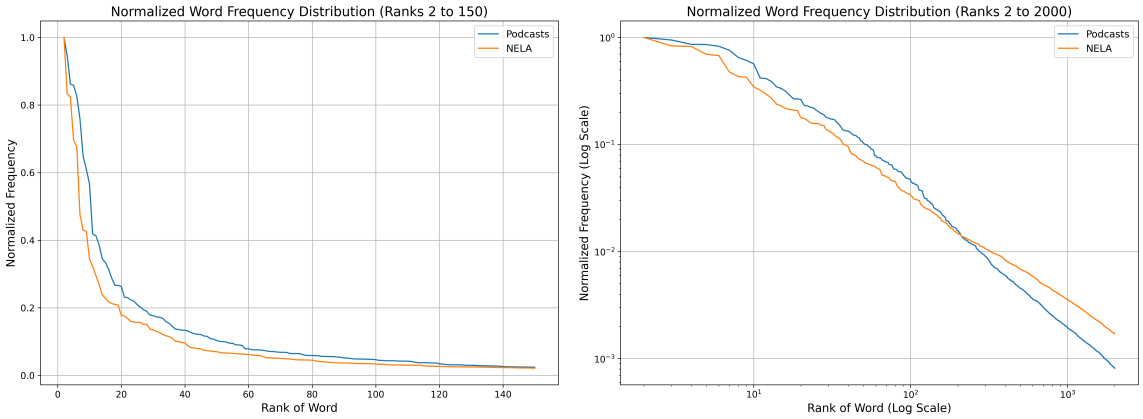

Fig. 7. Normalized word frequency distributions comparing Podcasts and NELA corpora.

In the left plot, the word frequency distribution for the podcast corpus falls off more gradually than that of the news corpus. This is also visible in the gap between the NELA and podcast slopes. It means that the top 200 words are used more evenly in the podcast corpus. However, in the right plot, right around word 200, the slopes cross and proceed to diverge slowly, where the NELA frequencies decline slower than their spoken word counterparts.

This finding is consistent with conversational speech containing a broader mix of common functional and pragmatic expressions [10]. Once beyond the top 200 items, human speech quickly falls back to a smaller set of recurring vocabulary. In contrast, the NELA curve's more gradual drop indicates that low- to mid-frequency words remain relatively more frequent in news articles. Written language maintains lexical diversity deeper into the tail. In summary, the distributions suggest that conversational speech has a flatter "head" of high-frequency words and a steeper "tail" of rare words, while written discourse has a sharper head but a richer, longer tail.

6.1.2 Isotropy & Intrinsic Dimensionality. Table 3 presents the isotropy of the embedding spaces across corpora and models. Isotropy measures the directional uniformity of an embedding space, which is to say that it checks whether word vectors, on average, point in similar or more different directions. High isotropy is generally considered good, as it means that vectors are well-spread in all directions rather than concentrated in a small space.

Isotropy was quantified by calculating its complement, anisotropy, and subtracting this value from 1. Anisotropy itself was measured as the average pairwise cosine similarity of sampled vectors. For each embedding model, a sample of 10,000 vectors was randomly selected; this sample size was chosen as preliminary tests showed it yielded stable isotropy scores. To ensure statistical robustness, this sampling and calculation process was repeated twenty times for each model. The mean (μ) and standard deviation (σ) of these twenty runs are reported in Table 3.

As stated, a higher score indicates greater uniformity and less directional bias in the embedding space. In theory, this should lead to more expressiveness, as vectors utilize more of the ambient space and thus have better capacity to encode distinct meanings [24, 50]. Because semantic directions are less correlated, highly isotropic embeddings usually separate an underlying corpus's features better than anisotropic ones.

The GloVe models clearly outperform the Word2Vec ones on this metric. Interestingly, the difference in isotropy across corpora is quite small per model, indicating that this metric is primarily influenced by model training rather than the underlying dataset. Even so, the difference in isotropy across corpora is somewhat more pronounced in

Model	μ	σ
Podcasts Word2Vec (avg)	0.51459	0.00254
NELA Word2Vec	0.58502	0.00199
Podcasts GloVe (avg)	0.92671	0.00136
NELA GloVe	0.93212	0.00144

Table 3. Isotropy analysis.

the Word2Vec models. This may suggest that the Word2Vec architecture is more sensitive to the characteristics of the underlying corpus than GloVe. The relationship between model architecture and corpus influence on embedding geometry could be an interesting avenue for future research.

To further analyze the geometric structure of the embedding space, Principal Component Analysis (PCA) was employed. PCA identifies the orthogonal directions, or principal components, that explain the maximum amount of variance in a dataset.

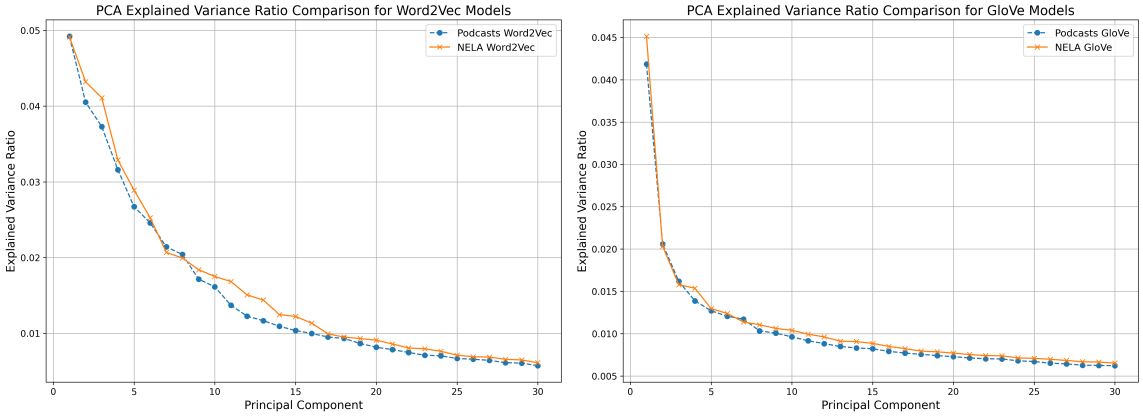

Fig. 8. PCA explained variance ratios for Word2Vec (left) and GloVe (right) embeddings.

Word2Vec and GloVe models were plotted separately, as principal components, like isotropy, were found to be more influenced by model architecture than by the underlying data.

These results are surprising. According to the isotropy analysis, GloVe-generated embedding spaces are more separated and well-spread in all directions, using more ambient space to encode distinct meanings than their Word2Vec counterparts. However, the PCA plots reveal the opposite trend: the explained variance for GloVe embeddings drops off much more sharply. The first component in the GloVe model explains around 4.3% of all variance, whereas the second merely accounts for 2%. In contrast, it takes the Word2Vec models 9 components to get below 2% of explained variance.

As a side note, within each model architecture, the slopes for both corpora are quite similar. This suggests that geometric properties like isotropy and principal component distribution are primarily artifacts of model architecture rather than corpus characteristics.

To further investigate the apparent contradiction between isotropy scores and PCA, I plotted the explained variance ratio of all 300 components (or dimensions) for all corpus-model combinations on a log-log scale in Figure 9. The results

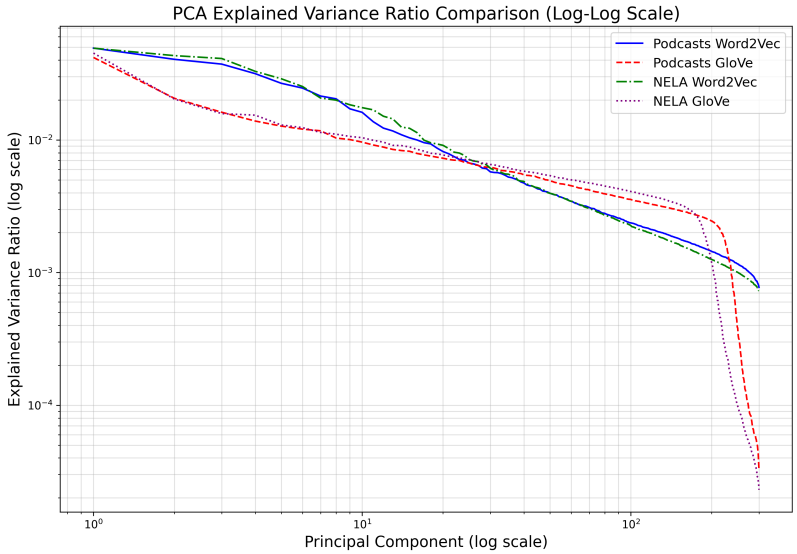

Fig. 9. PCA explained variance ratios for all embeddings and components.

are striking: the first 25 components of the Word2Vec embeddings explain much more variance than the GloVe ones. But then, from components 30 to 200, the trend reverses, and the GloVe model's explained variance declines much more slowly. The log-log plot makes it look like this difference is much less pronounced than the initial Word2Vec dominance, but in fact, for around 170 components, GloVe vectors explain around 10^{-3} more variance, which compounds.

Before continuing, I want to explain how high isotropy and a steep explained variance drop-off can be understood geometrically. Isotropy measures how uniformly distributed vectors are in an embedding space: highly isotropic spaces have vectors pointing in all directions from the origin. PCA, on the other hand, measures how well vectors are spread along different directions. In 3D, for instance, fully normalized vectors filling a sphere would maximize variance across all three dimensions. Yet, if only a few top components explain most of the variance, this indicates that the vectors largely lie on a lower-dimensional manifold.

We can reconcile these observations using a "pancake" analogy. Imagine the embedding vectors as forming a thin, flat pancake floating at the center of a much larger, empty sphere. The pancake lies along the directions of the top principal components. Some vectors reach toward the outer edge of the sphere along these principal directions, while others remain closer to the pancake's flat surface. Even though these shorter vectors do not extend far in magnitude, they still point in diverse directions across the pancake, preserving high isotropy. Thus, a vector space can simultaneously be highly isotropic and yet have most of its variance concentrated in a small number of principal components. This geometry is illustrated in Figure 10.

It follows that the GloVe models' embedding spaces might resemble a 300-dimensional equivalent of the topology in Figure 10. Visualizing how these embedding spaces evolve during training iterations presents an interesting avenue for future work.

Another notable feature in Figure 9 is that after the 200th component, both GloVe models fall off very sharply. I have two potential explanations for this. First, after the 200th component, the GloVe model may have had trouble extracting any new information from the corpus that wasn't already represented by a combination of its previous dimensions.

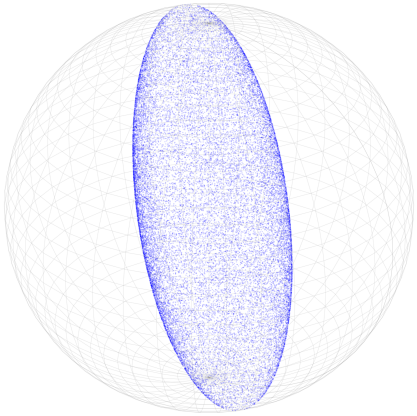

Fig. 10. What a highly isotropic, but low variance vector space may look like (in 3 dimensions).

This would mean that the last few dimensions of a vector only capture very specific nuances about the word they represent. I am somewhat skeptical of this explanation as I believe there are more than 200 dimensions along which one can define words, and the Word2Vec model seems to have no trouble filling out the last 100 dimensions. If GloVe regularly hit a ceiling at 200 dimensions, we'd probably see people using smaller GloVe models.

A second and more plausible explanation is that the GloVe models were not trained for a sufficient number of iterations to allow them to "fill in" all of their capacity for learning. The sharp drop-off pattern in both GloVe models (NELA and Podcasts) suggests this is architectural/training-related rather than corpus-specific. It is possible that GloVe's global optimization, which learns relationships across the entire vocabulary simultaneously, takes more iterations to fully utilize all 300 dimensions. In contrast, Word2Vec's local approach (predicting context words) might naturally fill out dimensions more evenly from early iterations since it is constantly making local predictions.

While the losses appeared to converge after 50 iterations, a review of the literature indicates that researchers regularly train GloVe models for 100-200+ iterations. It follows that retraining the models for more iterations is a clear extension of this work.

6.1.3 Qualitative Observations. A qualitative analysis of the top principal components was conducted to probe the primary semantic axes learned by the models. This involved examining the top 50 words most aligned with the positive and negative directions of each principal component. I focused on the Word2Vec embeddings, which displayed the most balanced and interpretable structure in their leading components (see Figure 8). Tables 4 and 5 reveal the findings for the first ten components.

The first observation from these tables challenges the common assumption that principal components in embedding spaces always represent clear semantic opposites. These dimensions do not necessarily represent a single concept distributed along a continuous axis, as might be expected. While PC1 seems to capture the formal-informal notion for both corpora (albeit flipped, potentially signaling a difference in priorities or communicative intent), subsequent dimensions don't span traditionally opposite concepts. This is not to say the concepts are semantically close; for example, *Food & Cooking* isn't particularly connected to *Disasters & Physical Events* (see Podcasts Word2Vec PC10).

PC	Podcasts Word2Vec	NELA Word2Vec
1	**Positive:** *Informal Language & Everyday Objects* – chili, sweater, taco, shorts, burger **Negative:** *Abstract & Academic Concepts* – governance, institutions, policies, economic, framework	**Positive:** *Politics, Law & Foreign Policy* – subpoenas, justice_department, legislation, sanctions, congress **Negative:** *Personal & Pop Culture* – lovely, birthday, fans, love_island, instagram
2	**Positive:** *Business & Technology* – manufacturers, software, startups, revenue, marketing **Negative:** *Religion & Archaic Language* – thee, thy, sinners, jesus_christ, repentance	**Positive:** *Economics & Industry* – manufacturers, efficiency, emissions, prices, supply **Negative:** *Social & Legal Controversies* – derek_chauvin, ahmaud_arbery, brett_kavanaugh, kneeling, acquitted
3	**Positive:** *Science, Health & Nutrition* – proteins, nutrients, cells, inflammation, hormones **Negative:** *Professional Titles & Media Roles* – nominated, senator, interviewed, hosted, attorney	**Positive:** *Local News & Pandemic Response* – curbside, curfew, hospitalizations, takeout, louisville **Negative:** *Abstract & Analytical Language* – qualities, arguably, critique, essence, undermines
4	**Positive:** *Podcast & Social Media Culture* – apple_podcasts, spotify, instagram, subscribe, patreon **Negative:** *News & Verbs of Action* – ruled, claimed, destroyed, reported, arrested	**Positive:** *Military & International Conflict* – artillery, troops, syria, missiles, invasion **Negative:** *Public Guidance & Services* – refunds, furlough, telehealth, guidelines, podcast
5	**Positive:** *Media & Character Analysis* – character, storytelling, compelling, cinematic, genre **Negative:** *Domestic & Religious Life* – groceries, unemployment, thy, passover, takeout	**Positive:** *Verbs of Action & Reporting* – announced, dispatched, returned, reported, confirmed **Negative:** *Theoretical & Scientific Concepts* – genes, viruses, capitalism, globalization, proteins

Table 4. Qualitative analysis of principal components (PC) 1-5 for Word2Vec models.

This finding is counter-intuitive and raises a key question: Are the models not strong enough to learn that opposites might be the best way to maximize variance across an embedding space?[4] Or maybe we aren't well-versed enough in our own language and don't have the analytical capacity to recognize that semantically, these poles actually do encode some form of cultural opposites that we can't consciously detect?

The most striking pattern is how differently each medium organizes crisis language. Both datasets document the same COVID surge, yet they semantically organized this shared experience into completely different structures. Podcasts integrated pandemic language with health/nutrition science (PC3: proteins, nutrients, inflammation), while news treated it as discrete crisis events (PC3,4: hospitalizations, guidelines, telehealth). This reveals how different media forms don't just report on crises differently; they literally think about them through different frameworks. The way scientific terms cluster differently across domains reinforces this: biological concepts in podcasts appear alongside lifestyle content, while in news they show up in outbreak contexts mentioning genes and viruses as global threats.

This cognitive difference also extends to political processing. The buildup to the 2020 election shows heavy political clustering in news embeddings (PC1), while podcasts focus on everyday culture (PC5). One interpretation is that podcasts, being more personal, tend to think of politics as it relates to individuals, semantically encoding political tension as cultural continuity instead of institutional preparation.

But what about the George Floyd protests and the broader Black Lives Matter uprising? Why don't they register as prominently in podcast embeddings as in news? Here, I propose two explanations. First, the transcription artifacts plaguing the podcast corpus likely fragmented proper names, particularly less common ones like Ahmaud Arbery.

[4]We could start encoding opposites into model training, thereby guiding the model to potential maximization of variance, but that would run counter to the generally accepted finding that the less human bias is put into the design of a system, the better the resulting model performs [13, 46]

PC	Podcasts Word2Vec	NELA Word2Vec
6	**Positive:** *Arts, Culture & History* – ancient, museum, architecture, paintings, mythology **Negative:** *American Sports & Athletes* – playoffs, quarterback, nfl, lakers, tom_brady	**Positive:** *Forensic & Scientific Investigation* – autopsy, dna, investigated, coroner, viruses **Negative:** *European Sports & Politics* – champions_league, premier_league, macron, brexit, manchester_united
7	**Positive:** *Identity & Health Issues* – allergies, transgender, hiv, racist, abortion **Negative:** *Religion & Sports* – tabernacle, touchdown, mvp, playoffs, saints	**Positive:** *Political Scandals & Tech* – burisma, mueller, fisa, bitcoin, iphone **Negative:** *Social & Public Health Issues* – starvation, poverty, deaths, homelessness, inequalities
8	**Positive:** *Interpersonal Interactions* – thanking, greeted, invited, smiled, grateful **Negative:** *Pop Culture & Political Ideology* – justice_league, socialism, batman, antifa, gameplay	**Positive:** *European News & Investigation* – nhs_england, wikileaks, assange, raab, portugal **Negative:** *US State Politics* – gop, wisconsin, california, republicans, democrats
9	**Positive:** *Geography, Culture & Sports* – northeast, cultures, olympic, traditions, soccer **Negative:** *Digital & Financial Administration* – refund, email, payment, password, lawsuit	**Positive:** *Public Health & Economic Projections* – hospitalizations, fauci, recession, forecast, fatalities **Negative:** *Rights & Legal Restrictions* – violates, prohibits, first_amendment, religious, laws
10	**Positive:** *Disasters & Physical Events* – ambulance, tornado, earthquake, helicopter, cameras **Negative:** *Food & Cooking* – delicious, bacon, cheese, garlic, vegetables	**Positive:** *European Football & Law* – arteta, champions_league, rosenstein, perjury, taser **Negative:** *Professions & Organizations* – founder, author, nonprofit, researcher, journalist

Table 5. Qualitative analysis of principal components (PC) 6-10 for Word2Vec models.

Because Whisper had limited training data for such names, it generated multiple variant spellings, diluting the semantic density of any single vector representation.[5]

Second, at a methodological level, the corpora embody different sampling philosophies. News outlets, despite their interpretive variations, tend to converge on shared event coverage: the NELA corpus was specifically curated to capture diverse perspectives on common stories during 2020's turbulence. Podcasts, by contrast, represent the entire publicly available ecosystem across genres, including health optimization, literary analysis, sports commentary, and spiritual guidance. This heterogeneity surfaces throughout the dimensional structure, with PC5's media analysis vocabulary suggesting robust literary subcultures within the dataset.

Furthermore, several other patterns specific to podcasts caught my attention. The heightened intimacy in PC8 (thanking, greeted, smiled) suggests podcasts embed more human emotional connection while news embeds more abstract cultural references. This likely reflects the conversational nature of podcasts: hosts and guests actually interacting creates semantic spaces filled with interpersonal warmth.

The religious emphasis throughout podcast embeddings is particularly interesting. Maybe some podcasters were drawing on spiritual frameworks to process contemporary events? Or maybe religious institutions turned to podcasts as a way to spread their message during COVID lockdowns? This finding is corroborated by observations from the data cleaning phase, which revealed a significant volume of religious content in the corpus.

Finally, there's the geographic divide. PC6 and PC8 reveal interesting geographic biases: podcast embeddings lean American (NFL, Lakers, Tom Brady) while news embeddings capture more European content (Premier League, Brexit, Macron). This suggests different media consumption patterns create different "mental maps" of what's globally important.

[5] I could have engineered corrective mappings to consolidate variant spellings into canonical forms, but consistency would have demanded similar treatment for all uncommon names throughout the corpus. Under time pressure, this was infeasible, but it remains a valid extension of this work.

It also reflects the fact that most English-speaking podcasts are US-centric, given the stark difference in European references.

In summary, this qualitative analysis shows how the same historical trauma; pandemic, protests and pre-election tension; crystallized into completely different semantic realities across media platforms. While podcasts maintained lifestyle-oriented, personal frameworks, news outlets developed specialized vocabularies for reporting on crises. More broadly, this deep dive reveals how identical historical moments create entirely different semantic worlds depending on media form. Exploring these findings further could reveal something important about how information ecosystems fragment shared experience into parallel meaning-worlds.

6.2 Downstream Task Performance

While linguistic and geometric evaluations offer valuable insights into the internal structure of word embedding spaces, their ultimate utility is determined by their performance on real-world NLP tasks. This section presents empirical findings from a comprehensive evaluation using intrinsic datasets compiled from multiple sources [58, 64], covering word similarity, analogy, categorization, and outlier detection tasks.[6]

6.2.1 Benchmarking. When referring to performance properties of either Podcasts Word2Vec or Podcasts GloVe in this subsection, I mean the average performance of Word2Vec and GloVe models across the 11-fold split of the dataset, respectively. All evaluations included coverage metrics to account for out-of-vocabulary words, with minimum thresholds applied to ensure reliable evaluation (e.g., minimum 3 words per cluster for outlier detection). These metrics (and more expansive results concerning this subsection, e.g., individual fold scores, standard deviations, and specific correct and false answers for data) can be found in my repository [53].

To start, word similarity tasks assess how well the geometric distance between word vectors correlates with human judgments of semantic similarity, measured by Spearman's rank correlation coefficient (ρ) between cosine similarity and human-assigned relatedness scores. Specifically, performance was calculated using Spearman's rank correlation between predicted cosine similarities and human judgments, with out-of-vocabulary words assigned zero vectors.

Model	mc30	men	mt287	mt771	rg65	rw	se17	sl999	sv3500	v143	ws353	wsr	wss	yp130
Podcasts Word2Vec (avg)	**68.6**	**72.6**	**69.2**	**65.5**	**71.2**	**42.6**	**67.3**	**28.4**	**31.3**	**53.6**	**70.5**	**67.1**	**73.2**	**54.5**
NELA Word2Vec	65.8	66.1	61.7	59.7	52.8	41.7	61.4	28.0	27.4	51.9	68.0	61.3	72.9	50.0
Podcasts GloVe (avg)	65.1	54.9	58.0	54.6	65.5	33.2	54.7	15.5	14.3	46.2	49.8	47.4	57.5	45.9
NELA GloVe	50.7	56.4	52.0	51.6	55.3	29.5	55.5	17.8	13.3	44.7	46.5	46.5	55.0	43.6

Table 6. Word similarity task performance (Spearman $\rho \times 100$).

Table 6 presents performance across 14 word similarity benchmarks.[7] Podcasts Word2Vec consistently outperforms other models, achieving the highest scores on all 14 datasets. Performance ranges from 28.4% on `sl999` to 73.2% on `wss`, with most scores spanning the high 50s to low 70s range. NELA Word2Vec shows consistently strong but lower performance across all benchmarks.

[6] To ensure compatibility with my (predominantly `gensim`-based) evaluation programs, I reformatted some of the pulled datasets. For others using `gensim` on intrinsic word embedding evaluation suites, I will upload them to my thesis repository [53].
[7] Evaluation set names for this (and other) tasks were abbreviated for readability, but, if required, can easily be inferred by scouring [64], which also contains short descriptions of each dataset (which are thus omitted here).

In contrast, GloVe models exhibit substantially weaker performance across both corpora, typically scoring 10 to 20 percentage points lower than their Word2Vec counterparts. The performance gap is consistent across diverse similarity benchmarks, suggesting a systematic advantage for Word2Vec's local context approach in capturing human-perceived semantic similarity over GloVe's global co-occurrence statistics.

The second set of benchmarks involved word analogy tasks. Analogy tasks probe vector space structure through dimensional relationships, solving analogies of the form "a is to b as c is to d" using vector arithmetic ($\vec{b} - \vec{a} + \vec{c}$). Intuitively, you take the point given by manipulating these three vectors, and then search for the closest vector to that point in the embedding space. Performance is measured by exact match accuracy, meaning only the top-ranked vector is considered a correct prediction.

Following this intuition, I implemented scoring such that analogies were solved using the standard `3CosAdd` method ($\vec{b} - \vec{a} + \vec{c}$), with top-1 accuracy calculated while excluding the input words from candidate predictions.

Model	google	jair	msr	sat	se2012
Podcasts Word2Vec (avg)	40.3	1.2	**47.9**	0.5	1.7
NELA Word2Vec	42.0	1.1	38.2	0.4	1.4
Podcasts GloVe (avg)	**44.4**	**3.5**	46.8	**1.3**	**2.2**
NELA GloVe	44.0	2.4	37.2	1.0	2.0

Table 7. Word analogy task performance (accuracy × 100).

The results in Table 7 reveal a reversal of the similarity task pattern. In contrast to similarity tasks, Podcasts GloVe achieves the highest performance on most benchmarks, scoring 44.4% accuracy on `google` analogies and 3.5% on `jair`. Our news-based GloVe model performs similarly well with 44.0% and 2.4% on those same datasets.

The notably low accuracy across `jair`, `sat`, and `se2012` datasets reflects their increased complexity compared to typical semantic analogies. The `jair` dataset (*Journal of Artificial Intelligence Research*) contains challenging analogies such as "Sounds is to echoes as waves is to [reflects]" and "Gas is to temperature as billiards is to [speeds]" that require domain-specific knowledge. Similarly, `sat` and `se2012` feature complex relationships like "Tunnel is to mine as conduit is to [fluid]" that extend beyond typical country-capital or syntactic analogies found in the `google` and `msr` datasets.

The analogy tasks also highlighted the downstream impact of preprocessing decisions. For instance, the choice made in section 3.2.3 to remove possessives had a direct, negative consequence for one benchmark. The `msr` dataset contains two analogy types, `NN_NNPOS` ("Mom is to mom's as dad is to [dad's]") and `NNPOS_NN`, that explicitly measure accuracy on possessives. Although this preprocessing step rendered the models unable to solve these specific analogies, they were still evaluated on the full dataset to ensure comparability with published results. This methodological detail underscores how initial data cleaning choices can have significant and unexpected downstream effects.

The third task, word categorization, evaluates an embedding's ability to group semantically similar words into coherent clusters. Clustering was performed using K-means [43] with the number of clusters set to match each dataset's ground truth categories. Performance was scored using two metrics: the Adjusted Rand Index (ARI) [33] and Normalized Mutual Information (NMI) [61].

Table 8 shows Word2Vec models generally outperforming GloVe across most datasets. Podcasts Word2Vec achieves the strongest performance on `bless` (ARI: 28.4, NMI: 61.9) and `bm` (ARI: 18.7, NMI: 44.5). NELA Word2Vec performs

Model	ap	bless	bm	essli
Podcasts Word2Vec (avg)	(13.5, **48.5**)	(**28.4**, **61.9**)	(**18.7**, **44.5**)	(15.1, 52.7)
NELA Word2Vec	(**15.8**, 46.4)	(22.2, 57.9)	(13.7, 41.3)	(10.5, 44.1)
Podcasts GloVe (avg)	(6.4, 39.8)	(14.6, 48.9)	(8.3, 34.9)	(22.0, 56.7)
NELA GloVe	(8.8, 45.5)	(12.9, 50.1)	(7.7, 34.6)	(**23.3**, **58.0**)

Table 8. Word categorization task performance (ARI, NMI × 100).

best on `ap` in terms of ARI (15.8), but is outperformed by its speech-to-text counterpart on the NMI metric (Podcasts Word2Vec scores 48.5).

The exception is `essli`, where NELA GloVe achieves the highest scores for both metrics (ARI: 23.3, NMI: 58.0). This suggests that global co-occurrence patterns may be beneficial for certain types of semantic categorization. Categorization has been the most balanced intrinsic task across model architectures so far.

Finally, let's look at outlier detection. Outlier detection identifies words that are semantically distant from others in a given set, typically by finding the word vector most distant from the group average.

Outlier scores were calculated using a hybrid approach combining centroid distance and average similarity to other words in the set. Specifically, outlier scores were computed as: `distance_to_centroid - average_similarity_to_peers`, with the highest-scoring words identified as outliers.

Model	888	ws500
Podcasts Word2Vec (avg)	78.7	**60.5**
NELA Word2Vec	**80.6**	59.7
Podcasts GloVe (avg)	68.5	48.5
NELA GloVe	72.5	47.2

Table 9. Outlier detection task performance (accuracy × 100).

Results in Table 9 demonstrate consistent Word2Vec superiority. NELA Word2Vec achieves 80.6% accuracy on the 8-8-8 dataset, while Podcasts Word2Vec leads on `ws500` with 60.5%. GloVe models perform substantially lower, with the best GloVe results being 72.5% on 8-8-8 (NELA GloVe) and 48.5% on `ws500` (Podcasts GloVe).

6.2.2 Scaling Effects. Figure 11 illustrates the impact of training data volume on downstream task performance, comparing the average scores of models trained on 11-fold splits of our SPoRC dataset (approximately 203 million words) to those trained on the full-scale 370,000 episode subset (approximately 2.2 billion words). Error bars for 11-fold splits are shown in gray.

Analysis of the plot reveals that all tasks benefit from increased training data, with performance improvements ranging from modest to substantial across all model-task combinations. This validates the general principle that larger corpora enable more robust word representations.

Both models demonstrate similar scaling benefits across tasks. For word similarity and categorization, the Word2Vec model improves marginally more, gaining 5.3% and 10.8% (compared to GloVe's 4.5% and 10.4%) respectively. Analogy improvements are similar (10.9% for Word2Vec and 11% for GloVe). GloVe improves a lot more in outlier detection (wins 5.2 percentage points), while scaling adds almost no performance for Word2Vec (2.3% increase).

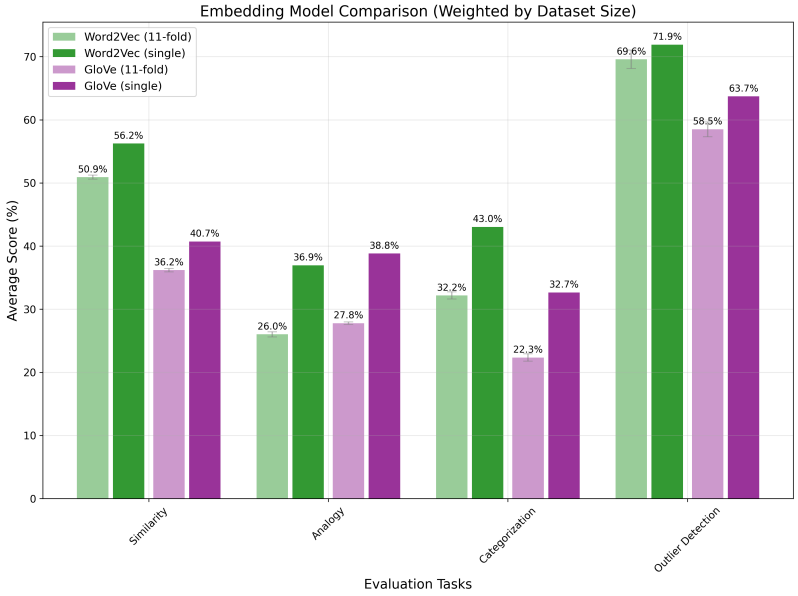

Fig. 11. Scaling's effects on task performance (SPoRC).

Among tasks, analogies show the most absolute improvement (almost 11%), closely followed by categorization. Outlier detection averages the least scaling gains across architectures.

6.3 Summary

First, an in-depth investigation of corpus linguistics and generated embeddings' vector space geometry provided fascinating insights:

- *Lexical Diversity*: The NELA (news) corpus has a higher TTR, indicating greater lexical diversity than the more naturally repetitive podcast corpus.
- *Word Frequencies*: Podcasts exhibit a flatter Zipfian "head" but a steeper "tail," suggesting a more even use of common words but less vocabulary depth than news.
- *Isotropy*: GloVe embeddings are significantly more isotropic than Word2Vec embeddings, a property that appears to be influenced more by a model's architecture than a corpus.
- *PCA Paradox*: GloVe's high isotropy contrasts with its steep PCA drop-off. This is explained by the "pancake analogy," where vectors are uniformly spread but largely confined to a low-dimensional manifold.
- *Media Revelations*: Where news embeddings focus on formal topics, podcast embeddings prioritize informal and personal subjects. And although both corpora chronicle the same events, they organize them into fundamentally different semantic frameworks, reflecting their unique communication styles.

And second, performance analysis on an extensive suite of intrinsic evaluation tasks revealed key patterns across models and datasets:

- *Complementary Strengths*: No single architecture proved universally optimal. While Word2Vec excels at similarity, categorization and outlier detection tasks, GloVe is superior on analogy tasks.

- *Corpus Characteristics*: Podcast-trained models generally outperform their news-based counterparts across most evaluation domains. Unsurprisingly, more diverse corpora seem to provide richer semantic representations.
- *Methodological Consequences*: The investigation of evaluation sets demonstrated the practical consequences preprocessing choices can have in NLP research. In this study, a tokenization decision rendered the word vectors unsuitable for a certain type of analogy.
- *Scaling Effects*: More training data consistently, albeit modestly, improved the SPoRC-based embeddings' performance, with some variance in impact (2-11%) depending on subtask and architecture.

6.3.1 A Note on Domain-Specificity. A critical methodological limitation emerged during analysis regarding the fundamental scope differences between the SPoRC and NELA-GT-2020 corpora. While temporal alignment and language calibration were maintained across both datasets, and despite evidence of overlapping event coverage (as demonstrated in this analysis and corroborated in [40]), the datasets exhibit fundamentally different topical boundaries.

NELA-GT-2020, despite encompassing diverse reliability and bias classifications, remains constrained to news content. Based on Gruppi et al.'s (2021) dataset description, the corpus predominantly consists of US, UK, and international news coverage. This represents a topically bounded sample of language use, focused primarily on current events, political discourse, and journalistic communication patterns.

Conversely, SPoRC encompasses transcriptions from all publicly available podcast episodes within the specified timeframe, resulting in substantially broader topical coverage. While podcasts frequently address current events and provide news analysis, the corpus additionally includes diverse domains such as sports commentary, religious discourse, culinary content, musical discussion, literary analysis, and scientific communication. This broader scope potentially provides more comprehensive coverage of natural language usage patterns and communicative contexts.

The observed superior performance of podcast-trained embeddings across evaluation tasks may therefore reflect not inherent advantages of the podcast medium for semantic representation, but rather the benefits of training on more topically diverse data that better approximates the full spectrum of human linguistic interaction.

This limitation does not invalidate the individual analyses presented, which remain methodologically sound and accurately reported. However, direct comparisons between news and podcast embedding performance should be interpreted with appropriate caution, as the observed differences may reflect domain coverage disparities rather than medium-specific linguistic characteristics. Future work might benefit from either topic-controlled comparisons or explicit investigation of domain diversity effects on embedding quality.

7 BIAS

This chapter investigates gender bias in SPoRC-based Word2Vec embeddings. This analysis builds on the methodology established in the landmark 2022 paper by Caliskan et al., *Gender bias in word embeddings* [15].

This gender bias analysis focuses specifically on the novel SPoRC corpus. As discussed in Chapter 2, extensive gender bias analysis has been done on news embeddings (e.g., the original Word2Vec paper [45] focused on the *Google News* corpus). Therefore, to contribute novel analysis to the field, this work focuses on the less-studied podcast corpus.

Gender associations and bias were generally calculated using SC-WEAT, as proposed by Caliskan et al. in 2017 [16]. A detailed explanation can be found in the background section (for even more details, see the original paper).

This chapter serves as a preliminary analysis demonstrating the research directions enabled by rigorously processed ASR transcript embeddings. The primary contribution is an initial investigation of gender bias within the novel SPoRC corpus. Our central hypothesis is that gendered language in a modern, conversational medium like podcasts differs from that in traditional written corpora, making it a worthy and under-studied subject. This analysis acts as a window into the societal associations present in this increasingly ubiquitous medium.

7.1 Preliminary Analysis

7.1.1 Frequent Words' Gender Associations. The first analysis examines the gender association of the embeddings representing the N most frequent words in the SPoRC corpus for both Word2Vec and GloVe models. Results are shown in Figure 12.

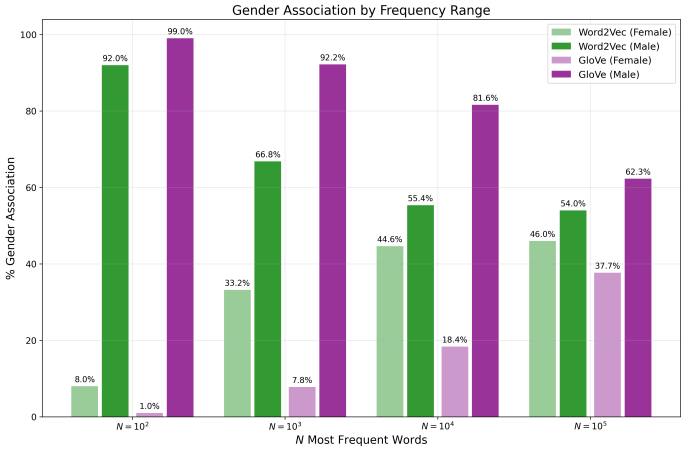

Fig. 12. Gender associations of top N most frequent corpus words for Word2Vec and GloVe models.

Both models' word embeddings show strong male associations in the first two categories, with Word2Vec and GloVe associating 92% and 99% of their 100 most frequent words with the male gender respectively.

The difference between the models is particularly striking, especially as N (the number of top most frequent words) grows. GloVe consistently associates much more of its frequent vocabulary with masculinity. It's more biased for $N = 10^3$ than Word2Vec is for $N = 10^2$, and the difference at $N = 10^4$ is striking (81.6% as opposed to 55.4% male-associated).

From the perspective of using embeddings as socioscientific diagnostic tools, this result is not necessarily negative; GloVe may simply be more sensitive to subtle biases within the corpus. However, for downstream applications such as CV-screening, such pronounced bias (especially in the GloVe model) would be highly problematic.

For clarity and focus, all subsequent analyses in this chapter are performed on the podcast-based Word2Vec embeddings. However, if interested, all code (and the embeddings) to run the corresponding analyses on SPoRC-based GloVe models (or even NELA models) can be found in the author's GitHub repository [53].

7.1.2 Gendered Clusters. Next, the geometric clustering of the 7,500 most female- and male-associated words was investigated. I created these using the notebooks provided by Ravfogel et al. for their INLP paper [56] (also why its title is *Original (t = 0)* – it's the clustering before the first iteration of their debiasing algorithm).

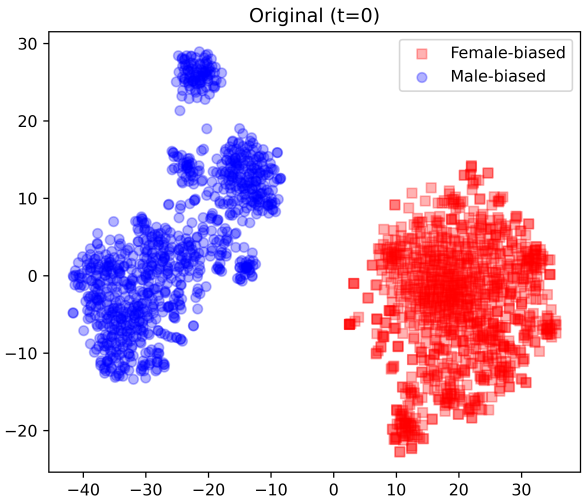

Fig. 13. Gendered clusters in the Word2Vec embedding space.

The clusters in Figure 13 exhibit a stark and clear separation. Interestingly, the male-biased cluster is somewhat segregated, with approximately three or four distinct subclusters. It would be interesting to investigate what semantic/syntactic/otherly-related word vectors these subclusters correspond to. The female-biased cluster is, by contrast, highly cohesive and forms a single, dense group, with, perhaps, traces of a subcluster forming on its southwestern shore.

Next, to investigate the semantic content of these clusters, a K-means analysis was performed on the top 1000 words for each model. To ensure the analyzed tokens were sufficiently represented in the corpus, I only included tokens that appeared in SPoRC at least 25 times for both female and male clusters.

I used the `scikit-learn`'s standard implementation of the K-means algorithm [51] (as introduced by [43]), and removed stopwords (using `nltk` [11]) in hopes of creating more meaningful semantic groups. The elbow plots for both male and female terms did not show clean elbows: the more clusters, the less inertia, with a quite constant downward

slope. The choice of $k = 11$ was selected as a trade-off between cluster granularity and conceptual clarity, and is a value also used in similar analyses by Caliskan et al. in [15].[8]

In part because the PCA analysis was so illuminating, I wanted to take apart the clusters, to reveal the semantic concepts that were most male- and female-biased. These are presented in Table 10.

Female clusters	**Male clusters**
Relationships & Family – relationships, marriage, dating, pregnant, woman	*Conflict & Combat* – war, fighting, kill, battle, enemies
Health & Wellness (Diet/Fitness) – diet, fitness, nutrition, lifestyle, vegan	*Strategy & Systems (Military/Finance)* – military, defense, stock, contract, numbers
Academia & Education – education, research, students, university, program	*Abstract Concepts & Judgment* – opportunity, trust, opinion, respect, problems
Female Names & Identities – girl, lady, queen, actress, sarah	*Organized Sports* – football, baseball, league, player, teams
Personal Growth & Wellness – journey, sharing, connect, experiences, meditation	*Competition & Achievement* – ahead, beat, chance, versus, pro
Home & Personal Aesthetics – beautiful, hair, fashion, cooking, kitchen	*Biblical Figures & Texts* – gospel, scriptures, paul, moses, priest
Public Spaces & Pandemic Life – lockdown, restaurant, shopping, masks, location	*Weapons & Hunting* – guns, weapons, shooting, hunting, sword
Business & Entrepreneurship – businesses, marketing, financial, clients, entrepreneur	*Professional Basketball (NBA)* – nba, jordan, lebron, finals, dunk
Digital Life & Social Media – online, facebook, instagram, website, netflix	*American Football (NFL)* – nfl, quarterback, draft, offense, touchdowns
Healthcare & Medicine – doctor, medical, hospital, therapy, nurse	*Christian Worship & Theology* – christ, holy, sin, salvation, worship
Arts & Literature – books, author, artist, stories, drama	*Physical Actions & The Outdoors* – throw, rock, fish, bear, mountain

Table 10. Qualitative analysis of semantically coherent clusters among the most gender-biased words in the embedding space.

Generally, these semantic clusters reinforce traditional gender stereotypes. They reveal that SPoRC (a recent and thus potentially more progressive dataset) contains gendered associations deeply embedded within it. Female-associated terms consistently cluster around traditionally feminine domains like relationships, health & wellness, home aesthetics, personal care, and caregiving roles. Male clusters, complementarily, gravitate towards traditionally masculine spheres, encompassing combat, organized sports, weapons, and competition in general. This suggests that podcasts, as a relatively new medium, may be amplifying existing gender roles found in legacy media rather than challenging them.

An embodiment of this phenomenon is the portrayed segregation in professional spaces. Female professional clusters include *Academia & Education* and *Healthcare & Medicine*, while male domains include financial and military strategy,

[8]The resulting clusters' 2-dimensional t-SNE [63] visualizations are omitted here as they lack obvious trends, but can be found in Figure 15 of the appendix.

religious leadership, and technical/engineering fields. However, the *Business & Entrepreneurship* cluster is strongly female-biased, which is noteworthy as this professional avenue is traditionally considered male-coded.[9]

Similarly, there seems to be a gender-based soft vs. hard skills divide. Female domains emphasize interpersonal connection, emotional intelligence, shared experiences, personal growth, and social relationship management.[10] In contrast, male domains focus on competition and physical action, a pattern likely influenced by the corpus's extensive sports commentary, but also abstract concepts and strategic thinking. These latter clusters are significant, as they appear to reinforce cultural narratives about emotional vs. rational capabilities being gender-linked.

Unsurprisingly, sports discourse is overwhelmingly male-coded. This is likely influenced by the US-centric nature of much of the podcast ecosystem. Sports discourse appears comprehensively male-coded across multiple distinct clusters, with, even, separate clusters emerging for individual leagues, like the NFL and NBA (next to just general organized sports and competitive achievement being male-associated too). There's a total absence of sports in female clusters. Thus, SPoRC embeddings probably perpetuate the marginalization of women's athletics.

A somewhat intriguing pattern emerges in technology-related clusters. *Digital Life & Social Media* appears prominently in female clusters. This association may reflect the rise of female influencers, content creators, and social media entrepreneurs in the podcasting ecosystem. Alternatively, it could indicate that social media is being framed as consumerist and relationship-based, which are concepts historically female-associated in news-based embeddings. But overall, this pattern suggests that podcasting may be creating new, and potentially more equitable, gendered narratives around digital engagement.

Finally, a comparison to the clusters identified in the Caliskan et al. study reveals that core stereotypical patterns show persistence across different media types.

Both datasets demonstrate similar clustering around beauty/appearance (female), sports (male), religious content (male), and health & relationships (female). However, small differences emerge in different tech spheres, with SPoRC-based embeddings female-associating life online, while Caliskan's analysis shows big-tech words to be strongly male-coded. Professional representation appears more sophisticated in podcast data, with female clusters including entrepreneurship categories absent in analyses of more traditional media.

Although not detailed in the table, an analysis of sexual content reveals it is also female-skewed in the SPoRC corpus, albeit perhaps at a lower and less degrading scale than observed in other corpora. This finding, along with Caliskan's analysis, suggests objectification patterns are persistent across media types.

While delivery mechanisms change, it seems that fundamental gender bias structures mostly persist. However, the emergence of a female-associated cluster for entrepreneurship indicates that the envisioned roles for women may be expanding or evolving within this medium, presenting a departure from patterns in more traditional datasets.

7.1.3 Bolukbasi-Kalai Stereotypes. To provide a qualitative assessment of SPoRC's embedded stereotypes, this section replicates a portion of the 2016 analogy-based analysis from Bolukbasi et al. [13].

Stereotypical associations in SPoRC were investigated by compiling a list of the most stereotypical male-female analogies found in the original *Man is to Computer Programmer as Woman is to Homemaker?* paper [13] and examining SPoRC's female analogies for each male term.

[9]This finding is significant, as such biases could have tangible effects on downstream applications, for instance by systematically disadvantaging certain candidates in professional contexts.

[10]Consistent with the personal development themes common in prominent female-hosted podcasts [1, 20, 22], this finding suggests the bias is a reflection of the source data rather than a model-induced artifact.

The analysis aims to uncover whether the podcast ecosystem reproduces, challenges, or creates new forms of gendered associations.

Male	Female	Podcasts #1	Podcasts #2-5
carpentry	sewing	nannying	joinery, woodshop, ceramics, metalwork
surgeon	nurse	surgeons	nurse, neurosurgeon, obstetrician, obgyn
burly	blond	mousy	gray_hair, buxom, matronly, voluptuous
chuckle	giggle	giggle	laugh, giggling, laughed, giggled
snappy	sassy	sassy	bitchy, snappier, spiced, snippy
football	volleyball	soccer	lacrosse, volleyball, netball, cheerleading
physician	registered nurse	obgyn	nurse, physicians, internist, practitioner
architect	interior designer	designer	architects, architecture, architectural, decorator
conservatism	feminism	feminism	progressivism, liberalism, feminist, feminists
guitarist	vocalist	singer	vocalist, drummer, bassist, songwriter
superstar	diva	supermodel	bianca_belair, grace_davis, ronda_rousey, becky_lynch
pizza	cupcakes	pizzas	pepperoni, calzone, pascuali, enchiladas
shopkeeper	housewife	lady	her, housemaid, receptionist, cashier
baseball	softball	softball	major_league_baseball, mlb, the_major_league_baseball, soccer
pharmaceuticals	cosmetics	pharmaceutical	medicines, medications, pharma, medication
lanky	petite	gawky	brunette, mousy, tomboyish, petite
affable	charming	vivacious	sassy, loquacious, charmingly, demure
brilliant	lovely	fabulous	lovely, fantastic, gorgeous, wonderful

Table 11. Top five female analogies in SPoRC Word2Vec embeddings.

The results are shown in Table 11. The columns *Male* and *Female* list the male and female parts of the stereotype found by Bolukbasi et al. in the original Google News Word2Vec embeddings. The *Podcasts #1* column shows the top-ranked female analogy for each male term. This analogy was generated by solving for d in the query "a is to b as c is to d," where $a = \overrightarrow{he}$, b is the vector for the male stereotype, and $c = \overrightarrow{she}$. The target vector d was calculated as $\vec{b} - \vec{a} + \vec{c}$, as described in Section 6.2. The *Podcasts #2-5* column lists the subsequent four most similar words.

A close examination of the results reveals several key patterns. In many cases, the SPoRC embeddings reproduce the exact stereotypes found in the original news corpus. For instance, the analogy for shopkeeper returns lady, and descriptive pairs like *burly:mousy*, *chuckle:giggle*, *snappy:sassy* and *brilliant:fabulous* are replicated (with varying degrees of precision). This suggests that many fundamental gender stereotypes are pervasive across different media types.

However, the analysis also reveals significant divergences, hinting at the unique linguistic environment of podcasts. For professional roles, the SPoRC embeddings occasionally produce less stereotypical analogies. The analogy for surgeon returns surgeons, not nurse, and the new pair *physician:obgyn* reflects a shift towards a peer profession rather than a subordinate one (though nurse appears in the top five for both *surgeon* and *physician*). *Architect* returns *designer*, which could be seen as a related profession rather than a subordinate one, but is likely an artifact of tokenization (the words *interior* and *designer*, if not recognized by NER, become separate tokens). Overall, this suggests that professional

associations in the podcast medium are evolving or, at a minimum, are represented differently than in traditional news text.

Furthermore, the embeddings surface associations uniquely characteristic of the podcast ecosystem's conversational and pop-culture-oriented nature. The top five analogies for superstar include the names of specific female celebrities and wrestlers (*bianca_belair*, *grace_davis*, *ronda_rousey*), a pattern not present in the original study. This finding indicates that the SPoRC embeddings capture a more informal, personality-driven semantic space. It also underscores the value of this corpus in reflecting contemporary cultural figures and linguistic patterns that are often absent in more formal, edited texts.

7.2 Summary

This chapter analyzes gender bias in SPoRC podcast embeddings, revealing both the persistence of traditional stereotypes and the emergence of novel associations unique to the medium. While concepts like family and home remain female-coded and themes of conflict and sports male-coded, we also observe significant shifts. These include the female-association of business and entrepreneurship, less hierarchical professional analogies (e.g., *physician* to *obgyn*), and the appearance of contemporary pop-culture figures. The findings suggest that while core biases are resilient, the podcast ecosystem reflects and potentially shapes evolving gender roles, underscoring its value as a subject for sociolinguistic study.

8 DISCUSSION

This thesis embarked on an investigation into the challenges and characteristics of word embeddings derived from large-scale, ASR-transcribed corpora. The initial goal of analyzing gender bias in podcasts evolved into a foundational, data-centric inquiry upon discovering the profound impact of the transcription process itself on data integrity and semantic structure. The findings carry significant implications for NLP research, computational social science, and the development of more robust language technologies. This chapter synthesizes these implications, acknowledges the study's limitations, and outlines promising directions for future work.

8.1 Implications

The results of this work underscore that ASR-transcribed corpora are not merely larger, noisier versions of written text; they are distinct data modalities with unique artifacts that can fundamentally alter empirical findings [3, 48].

For NLP Research and Practice. The most critical practical contribution is the identification and mitigation of looping hallucinations in Whisper-transcribed audio [39]. While these artifacts constitute a small fraction of the total corpus by character count (0.959%) , their qualitative impact is severe, as they can overwrite and erase genuine spoken content, representing a significant form of data corruption [36]. The novel, hash-based methodology developed in this thesis provides a scalable and reproducible framework for detecting and filtering such errors, offering a crucial tool for researchers working with similar datasets.

Furthermore, this work serves as a case study in the necessity of rigorous, iterative preprocessing. The discovery of undocumented artifacts in both the ASR-transcribed podcast corpus and the web-scraped news corpus, such as non-speech audio markers in SPoRC and foreign-language content in NELA-GT-2020, highlights that no large-scale dataset can be treated as "clean" out of the box [9, 21]. Default NLP pipelines may be insufficient for handling the diversity of noise present in real-world data. Additionally, the scalable, parallelized preprocessing architecture developed here demonstrates computational efficiency gains that could benefit researchers working with similar large-scale datasets.

For Computational Social Science. A central finding is that different media construct divergent "semantic realities" from the same shared historical events [17, 30]. During the tumultuous events of 2020, news media framed crises around institutional language, politics, and law, while podcasts prioritized personal lifestyle, health, and interpersonal dynamics. This demonstrates that the choice of corpus is not a neutral act; the medium itself imposes a semantic framework that shapes how societal meaning is encoded and, consequently, what conclusions can be drawn from the resulting embeddings. This finding has methodological implications for computational sociolinguistics, suggesting that researchers must carefully consider not just the content but the communicative medium of their corpora when making claims about societal language patterns.

The gender bias analysis further complicates our understanding of language and social change. While the SPoRC embeddings perpetuate many traditional gender stereotypes seen in legacy media; associating female-coded terms with family and home, and male-coded terms with conflict and sports; they also surface novel patterns. The strong female association with "Business & Entrepreneurship" (see Table 10) and the shift in professional analogies (e.g., *physician:obgyn* instead of *physician:nurse*) suggest that the conversational and diverse nature of podcasts may reflect and shape evolving social roles [12, 13, 16]. This positions the medium as a valuable, albeit complex, resource for studying cultural shifts.

8.2 Limitations

While this study provides novel methods and insights, its conclusions should be considered within the context of several limitations.

The most significant limitation is the topical disparity between the SPoRC and NELA-GT-2020 corpora. SPoRC encompasses a vast range of genres, while NELA-GT-2020 is constrained to news content. The observed superior performance of podcast-trained embeddings on many downstream tasks may be an effect of this broader topical coverage rather than an inherent property of spoken language.

Methodologically, the analysis is also bounded in several ways. First, the hallucination analysis is specific to OpenAI's Whisper and the observed error patterns may not generalize to other ASR systems [3, 37]. Second, preprocessing involved necessary trade-offs with downstream consequences; for instance, the decision to remove possessives negatively impacted performance on specific analogy benchmarks.

Finally, the evaluation approach relied primarily on intrinsic tasks rather than extrinsic applications, which limits our ability to assess real-world performance in tasks like sentiment analysis or named entity recognition.

8.3 Extensions & Future Work

The findings and limitations of this thesis open several promising avenues for future research.

Technical and Methodological Extensions. Future work should apply the hallucination detection pipeline to transcripts from other ASR providers (e.g., Google Speech-to-Text, Azure) to create a comparative taxonomy of ASR artifacts. This methodology could also be adapted from a post-hoc cleaning script into a real-time filter to improve the quality of streaming ASR applications. To build upon the intrinsic evaluations, the cleaned embeddings should be tested on a range of extrinsic, downstream tasks to assess their practical utility in real-world scenarios.

Sociolinguistic and Bias-focused Research. A longitudinal analysis of podcast corpora spanning several years could track the evolution of semantic frameworks and social biases over time, offering insights into the pace of cultural change. Extending this work to non-English and non-US-centric podcasts would be crucial for understanding how ASR artifacts and media-specific biases manifest across different linguistic and cultural contexts. Cross-media comparative studies could systematically examine how the same events and topics are semantically organized across different media types (podcasts, news, social media), providing deeper insights into medium-specific semantic construction. Building on the preliminary debiasing exploration in the appendix, a more rigorous study of debiasing techniques is needed, one that evaluates their impact on fairness in extrinsic applications, ensuring they truly mitigate harm rather than simply obscuring it [15, 44].

The serendipitous discovery of looping hallucinations, which resulted from an accidental file-splicing error, serves as a final, crucial lesson. It highlights the indispensable value of qualitative, manual data inspection in an era of large-scale, automated analysis. Future research must continue to bridge quantitative rigor with qualitative curiosity to navigate the complex, artifact-ridden, and semantically rich world of modern language data.

9 CONCLUSION

This thesis was initially conceived to analyze gender bias in podcasts. However, a preliminary investigation revealed that the foundational challenge lay in the integrity of the ASR-transcribed data itself. The research, therefore, pivoted to address this core issue.

The primary contribution of this work is a novel methodological framework for processing and sanitizing large-scale spoken-language corpora. This framework includes a scalable preprocessing pipeline and a robust, database-driven system for systematically detecting and quantifying 'looping hallucinations,' a severe and previously under-studied form of data corruption present in Whisper-generated transcripts.

This foundational work unlocked a key insight: from the same historical events, news and podcast corpora construct divergent "semantic realities." Where news media framed crises through an institutional lens, podcasts built a world of personal and lifestyle-oriented narratives. The analysis of social dynamics revealed a similar tension: while the podcast medium reproduces many traditional gender stereotypes, it also reflects evolving social roles, underscoring the value of ASR corpora for nuanced sociolinguistic study.

In conclusion, this thesis demonstrates the critical importance of foundational, data-centric research in an era of large-scale, automated analysis. The discovery of looping hallucinations, an artifact constituting less than 1% of the corpus but capable of significant data corruption, underscores the necessity of combining quantitative methods with rigorous data inspection. By developing and applying a methodology to address such hidden complexities, this work provides a more reliable foundation for future research into the rich and expanding domain of spoken-language data.

ACKNOWLEDGMENTS

First and foremost, I would like to express my sincere gratitude to my supervisor, Dr. Alexander Hoyle, for his invaluable guidance and mentorship throughout this project. His sharp insights and consistent feedback were instrumental in shaping this thesis and ensuring it met a high academic standard. Thank you for enlightening me that a *Motivation* section isn't about personal motivation, and a thesis isn't a blog post. I am also deeply grateful to Professor Ryan Cotterell for graciously agreeing to supervise this work.

This research would not have been possible without the foundational work of others. I extend my thanks to Ben Litterer, David Jurgens, and Dallas Card for their effort in compiling the SPoRC dataset, and to Mauricio Gruppi for granting me access to the NELA-GT-2020 corpus.

I was fortunate to have the opportunity to collaborate with the Rycolab, and I thank Niklas Stöhr and the entire group for welcoming me. My initial introduction to the lab came from Connor Charlton, who suggested I look into writing my thesis there, a recommendation for which I am very thankful.

A special note of appreciation goes to my friends and colleagues. To Carl Dahlberg, thank you for the memorable and productive writing camp on the French coast. To my friends at ETH, thank you for the stimulating conversations, fresh perspectives, and constant encouragement that kept me going.

Finally, I owe deep thanks to my parents. Thank you for your unwavering support, encouragement, and for meticulously proofreading my drafts. This would not have been possible without you.

REFERENCES

[1] 2025. 80 Best Personal Development & Self Improvement Podcasts For Women. https://podcast.feedspot.com/personal_development_podcasts_for_women/. Accessed: 2025-08-23.

[2] Eneko Agirre, Enrique Alfonseca, Keith Hall, Jana Kravalova, Marius Pasca, and Aitor Soroa. 2009. A study on similarity and relatedness using distributional and WordNet-based approaches. In *Proceedings of the NAACL HLT 2009 Workshop on Computational Linguistics for Linguistic Complexity*. 19–27.

[3] Terry Amorese, Claudia Greco, Marialucia Cuciniello, Rosa Milo, Olga Sheveleva, and Neil Glackin. 2023. Automatic speech recognition (ASR) with Whisper: Testing performances in different languages. In *S3C@CHItaly*. 1–8.

[4] Maria Antoniak and David Mimno. 2021. Bad seeds: Evaluating lexical methods for bias measurement. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. 1889–1904.

[5] Carlos Arriaga, Alejandro Pozo, Javier Conde, and Alvaro Alonso. 2024. Evaluation of real-time transcriptions using end-to-end ASR models. In *arXiv preprint arXiv:2409.05674*.

[6] Mateusz Barański, Jan Jasiński, Julitta Bartolewska, Stanisław Kacprzak, Marcin Witkowski, and Konrad Kowalczyk. 2025. Investigation of whisper ASR hallucinations induced by non-speech audio. In *ICASSP 2025 IEEE International Conference on Acoustics, Speech and Signal Processing*. 1–5.

[7] Marco Baroni, Georgiana Dinu, and Germán Kruszewski. 2024. Don't count, predict! A systematic comparison of context-counting vs. context-predicting semantic vectors. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 238–247.

[8] Stefan Behnel, Robert Bradshaw, Craig Citro, Lisandro Dalcin, Dag Sverre Seljebotn, and Kurt Smith. 2011. Cython: The best of both worlds. In *SciPy*, Vol. 17. Austin, TX, 29.

[9] Emily M. Bender and Batya Friedman. 2018. Data Statements for Natural Language Processing: Toward Mitigating System Bias and Enabling Better Science. *Transactions of the Association for Computational Linguistics* 6 (2018), 587–604.

[10] Douglas Biber. 1986. Spoken and written textual dimensions in English: Resolving the contradictory findings. *Language* (1986), 384–414.

[11] Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural Language Processing with Python: Analyzing Text with the Natural Language Toolkit*. O'Reilly Media, Inc.

[12] Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. 2020. Language (technology) is power: A critical survey of "bias" in NLP. *arXiv preprint arXiv:2005.14050* (2020).

[13] Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. 2016. Man is to computer programmer as woman is to homemaker? Debiasing word embeddings. *Advances in Neural Information Processing Systems* 29 (2016).

[14] Hervé Bredin. 2023. pyannote.audio 2.1 speaker diarization pipeline: principle, benchmark, and recipe. In *24th INTERSPEECH Conference (INTERSPEECH 2023)*. ISCA, 1983–1987.

[15] Aylin Caliskan, Pimparkar Parth Ajay, Tessa Charlesworth, Robert Wolfe, and Mahzarin R Banaji. 2022. Gender bias in word embeddings: A comprehensive analysis of frequency, syntax, and semantics. In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*. 156–170.

[16] Aylin Caliskan, Joanna J Bryson, and Arvind Narayanan. 2017. Semantics derived automatically from language corpora contain human-like biases. *Science* 356, 6334 (2017), 183–186.

[17] Wallace L Chafe and Deborah Tannen. 1987. The relation between written and spoken language. *Annual Review of Anthropology* 16, 1 (1987), 383–407.

[18] Ann Clifton, Sravana Reddy, Yongze Yu, Aasish Pappu, Rezvaneh Rezapour, Hamed Bonab, Maria Eskevich, Gareth Jones, Jussi Karlgren, Ben Carterette, et al. 2020. 100,000 podcasts: A spoken English document corpus. In *Proceedings of the 28th International Conference on Computational Linguistics*. 5903–5917.

[19] Scott Deerwester, Susan T Dumais, George W Furnas, Thomas K Landauer, and Richard Harshman. 1990. Indexing by latent semantic analysis. *Journal of the American Society for Information Science* 41, 6 (1990), 391–407.

[20] Amy Diehl and Leanne Dzubinski. 2024. Glass Walls: Shattering the Six Gender Bias Barriers Still Holding Women Back at Work. https://thestoryofwomanpodcast.com/episode/s3-e4-shattering-the-six-gender-bias-barriers-still-holding-women-back-at-work-with-amy-diehl-and-leanne-dzubinski-authors-of-glass-walls. Accessed: 2025-08-23.

[21] Jesse Dodge, Maarten Sap, Ana Marasović, William Agnew, Gabriel Ilharco, Dirk Groeneveld, Margaret Mitchell, and Matt Gardner. 2021. Documenting large webtext corpora: A case study on the colossal clean crawled corpus. *arXiv preprint arXiv:2104.08758* (2021).

[22] Catherine D'Ignazio and Lauren Klein. 2021. Data Feminism and Cultural Biases in Data. https://www.preposterousuniverse.com/podcast/2021/07/19/156-catherine-dignazio-on-data-objectivity-and-bias/. Accessed: 2025-08-23.

[23] Edison Research and Triton Digital. 2024. The Infinite Dial 2024. https://www.edisonresearch.com/the-infinite-dial-2024/

[24] Kawin Ethayarajh. 2019. How contextual are contextualized word representations? Comparing the geometry of BERT, ELMo, and GPT-2 embeddings. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. 55–65.

[25] Lev Finkelstein, Evgeniy Gabrilovich, Yossi Matias, Ehud Rivlin, Zach Solan, Gadi Wolfman, and Eytan Ruppin. 2001. Placing search in context: The concept revisited. In *WWW*. 406–414.

[26] Rita Frieske and Bertram E Shi. 2024. Hallucinations in neural automatic speech recognition: Identifying errors and hallucinatory models. *arXiv preprint arXiv:2401.01572* (2024).

[27] Sharon Goldwater, Daniel Jurafsky, and Christopher D Manning. 2010. Contextual appropriateness: Describing and detecting unscripted ASR errors. In *Proceedings of the NAACL HLT 2010 Workshop on Speech and Language Processing for Assistive Technologies*. 74–82.

[28] Hila Gonen and Yoav Goldberg. 2019. Lipstick on a pig: Debiasing methods cover up systematic gender biases in word embeddings but do not remove them. *arXiv preprint arXiv:1903.03862* (2019).

[29] Maurício Gruppi, Benjamin D Horne, and Sibel Adalı. 2021. NELA-GT-2020: A large multi-labelled news dataset for the study of misinformation in news articles. *arXiv preprint arXiv:2102.04567* (2021).

[30] M A K Halliday. 1989. *Spoken and Written Language*. Oxford University Press.

[31] Zellig S Harris. 1954. Distributional structure. *Word* 10, 2-3 (1954), 146–162.

[32] Amber Holder and Chris Lygate. 2021. Podcasting as the New Oral History: Media, Storytelling and Community Engagement. *Oral History* 49, 2 (2021), 5–17.

[33] Lawrence Hubert and Phipps Arabie. 1985. Comparing partitions. *Journal of Classification* 2, 1 (1985), 193–218.

[34] Masahiro Kaneko and Danushka Bollegala. 2019. Gender-preserving debiasing for pre-trained word embeddings. *arXiv preprint arXiv:1906.00742* (2019).

[35] Os Keyes. 2017. *Stop Mapping Names to Gender*. https://ironholds.org/names-gender/

[36] Allison Koenecke, Anna Seo Gyeong Choi, Katelyn X Mei, Hilke Schellmann, and Mona Sloane. 2024. Careless whisper: Speech-to-text hallucination harms. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*. 1672–1681.

[37] Korbinian Kuhn, Verena Kersken, Benedikt Reuter, Niklas Egger, and Gottfried Zimmermann. 2024. Measuring the accuracy of automatic speech recognition solutions. *ACM Transactions on Accessible Computing* 16, 4 (2024), 1–23.

[38] Brian N Larson. 2017. Gender as a variable in natural-language processing: Ethical considerations. Association for Computational Linguistics.

[39] Fangzhou Li, Hao Wang, Lin Lin, et al. 2025. Demystifying Hallucination in Speech Foundation Models. In *Findings of ACL 2025*. https://aclanthology.org/2025.findings-acl.1190/

[40] Benjamin Litterer, David Jurgens, and Dallas Card. 2024. Mapping the Podcast Ecosystem with the Structured Podcast Research Corpus. *arXiv preprint arXiv:2411.07892* (2024).

[41] Dario Llinares and Stella Hockenhull. 2023. *Podcast Studies: Practice into Theory*. Routledge.

[42] Reza Lotfian and Carlos Busso. 2017. Building naturalistic emotionally balanced speech corpus by retrieving emotional speech from existing podcast recordings. *IEEE Transactions on Affective Computing* 10, 4 (2017), 471–483.

[43] J MacQueen. 1967. Multivariate observations. In *Proceedings of the 5th Berkeley Symposium on Mathematical Statistics and Probability*, Vol. 1. 281–297.

[44] Nicholas Meade, Elinor Poole-Dayan, and Siva Reddy. 2021. An empirical survey of the effectiveness of debiasing techniques for pre-trained language models. *arXiv preprint arXiv:2110.08527* (2021).

[45] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781* (2013).

[46] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. *Advances in Neural Information Processing Systems* 26 (2013).

[47] James Milton. 2009. *Measuring second language vocabulary acquisition*. Vol. 45. Multilingual Matters.

[48] Luke K Miner, Tamer Transaction, Peter D Harms, and Dustin Wood. 2020. Assessing the golden standard: a meta-analysis of 35 years of M Turk. *Journal of Applied Psychology* 105, 10 (2020), 1136.

[49] Theodora Moldovan. 2025. Exploring Collective Action in Black Lives Matter Podcast Conversations.

[50] Jiang Mu and Prateek Viswanath. 2018. All-but-the-Top: Simple and Effective Postprocessing for Word Representations. In *International Conference on Learning Representations (ICLR)*.

[51] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* 12 (2011), 2825–2830.

[52] Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. GloVe: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*. 1532–1543.

[53] Jonathan Quinn. 2025. Bachelor's Thesis Repository. https://github.com/jofras/bachelors-thesis. Accessed: 2025-08-23.

[54] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *International conference on machine learning*. PMLR, 28492–28518.

[55] Shauli Ravfogel. 2020. Iterative Nullspace Projection. https://github.com/shauli-ravfogel/nullspace_projection.

[56] Shauli Ravfogel, Yanai Elazar, Hila Gonen, Michael Twiton, and Yoav Goldberg. 2020. Null it out: Guarding protected attributes by iterative nullspace projection. *arXiv preprint arXiv:2004.07667* (2020).

[57] Gerard Salton and Michael J McGill. 1983. *Introduction to modern information retrieval*. McGraw-Hill, Inc.

[58] Tobias Schnabel, Igor Labutov, David Mimno, and Thorsten Joachims. 2015. Evaluation methods for unsupervised word embeddings. In *Proceedings of the 2015 conference on empirical methods in natural language processing*. 298–307.

[59] Surabhi Shripal Singh. 2023. Recommendation Framework on English Speaking Podcasts using Textual Information Analysis. *Dublin, National College of Ireland* (2023).

[60] Stanford NLP Group. [n. d.]. GloVe GitHub Repository. https://github.com/stanfordnlp/GloVe. Accessed: 2025-03-13.

[61] Alexander Strehl and Joydeep Ghosh. 2002. Cluster ensembles—a knowledge reuse framework for combining multiple partitions. In *Journal of machine learning research*, Vol. 3. 583–617.

[62] Francisco B Valero, Marion Baranes, and Elena V Epure. 2022. Topic modeling on podcast short-text metadata. In *European Conference on Information Retrieval*. Springer, 472–486.

[63] Laurens van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-SNE. *Journal of Machine Learning Research* 9, Nov (2008), 2579–2605.

[64] Vecto-AI. [n. d.]. word-benchmarks: Benchmarks for intrinsic word embedding evaluation. https://github.com/vecto-ai/word-benchmarks. Accessed: 2025-05-20.

[65] Jieyu Zhao, Yichao Zhou, Zeyu Li, Wei Wang, and Kai-Wei Chang. 2018. Learning gender-neutral word embeddings. *arXiv preprint arXiv:1809.01496* (2018).

[66] George Kingsley Zipf. 1949. *Human Behavior and the Principle of Least Effort*. Addison-Wesley, Cambridge, MA.

[67] Radim Řehůřek and Petr Sojka. 2010. Gensim: Topic Modelling for Humans. https://radimrehurek.com/gensim/ Version 4.0.0.

A MODEL TRAINING AND CONFIGURATION

The specific CPU and memory configurations used on the HPC cluster, along with the final wall-clock training times and loss curves for each model, are presented in Table 12 and Figure 14.

Model	Cores used	Memory per CPU (GB)	Wall-clock time
Podcasts Word2Vec	64	2	4h 21m 24s
NELA Word2Vec	64	1	19m 38s
Podcasts GloVe	32	4	3h 57m 43s
NELA GloVe	32	2	2h 57m 14s

Table 12. Model training times and computational resources used on the *Euler* cluster.

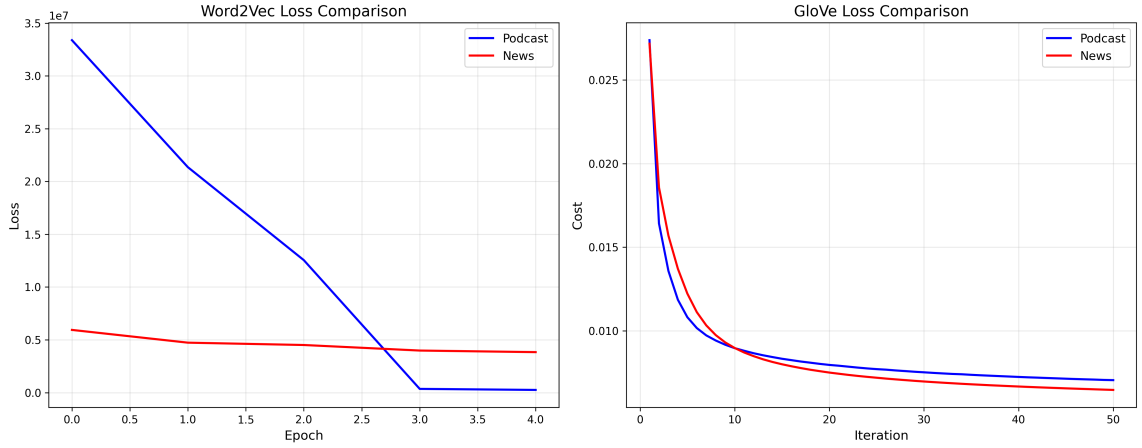

Fig. 14. Training loss comparisons across models and corpora.

B SUPPLEMENTARY VISUALIZATIONS

The t-SNE visualizations of the top 1,000 most gender-associated terms were omitted from the main analysis to avoid potential over-interpretation. The t-SNE algorithm is highly sensitive to hyperparameter tuning, and its output can be misleading if not carefully calibrated. The plots in Figure 15 were generated using standard parameters for exploratory purposes and should be interpreted with this limitation in mind.

A visual inspection of the plots reveals distinct structural differences between the male- and female-associated clusters. The female-associated clusters, for example, appear more dispersed across the coordinate space. In contrast, the male-biased clusters form a more elongated, contiguous structure.

C EXPLORATORY DEBIASING ANALYSIS

This section details a preliminary analysis of debiasing the SPoRC embeddings using the Iterative Nullspace Projection (INLP) algorithm [56]. As the primary focus of this thesis is on the analysis of inherent corpus biases rather than their mitigation, this exploration was moved from the main text. It is included here for completeness and to document

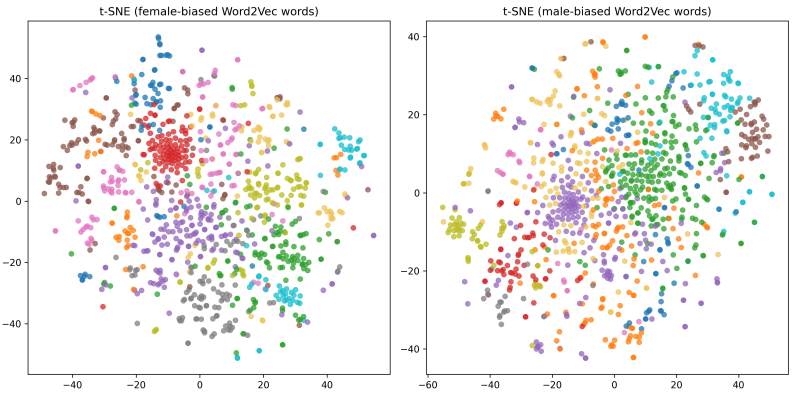

Fig. 15. t-SNE of clusters of top 1000 most female- and male-biased SPoRC Word2Vec words.

initial findings. This analysis suggests several avenues for future work, such as a topological analysis of the debiased embedding space or a study of which specific embeddings are most affected by the process.

C.1 Debiasing Procedure

The podcast Word2Vec embeddings were debiased using Iterative Nullspace Projection (INLP), an algorithm introduced by Ravfogel et al. (2020) [56]. The implementation from the original authors' publicly available code repository [55] was used.

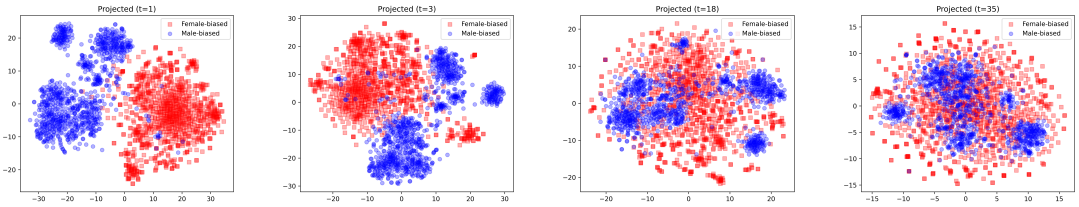

Fig. 16. Embedding space debiasing across INLP iterations.

The visual effect of applying 35 iterations of INLP is displayed in Figure 16. Comparing these plots to the original clustering in Figure 13, even a single iteration brings the gendered clusters significantly closer. The clusters begin to mix substantially in the later iterations, indicating a reduction in the geometric separability of gender-associated words.

Interestingly, the subclusters of male-biased terms remain somewhat disjoint even after the final iteration. Further iterations may be required to achieve full mixing of these subclusters. Female-biased terms, in contrast, spread more evenly after the algorithm is applied.

C.2 Impact on Downstream Performance

To estimate the effect of debiasing on downstream performance, the evaluation suite from Chapter 6 was run on the debiased embeddings. The results are presented in Table 13.

Task	Raw	Debiased
Similarity	0.5625	**0.5635**
Analogy	**0.3695**	0.3339
Categorization	**0.4303**	0.3508
Outlier Detection	**0.7191**	0.7102

Table 13. Comparing podcast Word2Vec performance pre- and post-debiasing.

Overall, debiasing degraded embedding performance on most intrinsic evaluation tasks. While performance on similarity and outlier detection tasks remained relatively stable, both analogy and categorization tasks suffered significant losses (3.56 and 7.95 percentage points, respectively).

The performance degradation on categorization tasks is particularly noteworthy. It may suggest that geometric directions associated with gender are also leveraged by the model for ostensibly non-gendered categorization. Alternatively, it could indicate that the evaluation datasets themselves contain latent gender biases that align with those in the original embeddings.

C.3 Impact on Analogical Stereotypes

To provide a qualitative assessment of the debiasing algorithm's effectiveness, the analogy-based stereotype analysis from Bolukbasi et al. (2016) [13] was replicated on both the original and debiased embeddings.

The analysis of the debiased analogies, shown in Table 14, reveals several key patterns.

Positive Effects. Debiasing was effective for many professional stereotypes. For instance, the female analogy for *carpentry* shifted from the stereotypical *nannying* to the topically relevant *woodworking*. Similarly, *physician*'s analogy shifted from *obgyn* to the more parallel *physicians*.

Arbitrary Shifts and Semantic Displacement. In other cases, INLP appears to displace bias rather than eliminate it. For sports analogies, it produced arbitrary changes (e.g., mapping both *football* and *baseball* to *basketball*) with some semantic loss. A more striking example is *snappy*, whose female counterpart shifted from *sassy* to *horseradish*, indicating that the algorithm reinterpreted the word's meaning from a behavioral trait to a taste profile. The emergence of specific proper names (e.g., *katherine_nikolai*, *derek_green*) suggests the system latches onto incidental corpus entities when gendered reasoning paths are removed.

Semantic Neutralization. The most revealing failures occur where debiasing produces outputs that are technically gender-neutral but semantically impoverished. For example, shifting the analogy for *brilliant* from *lovely* to *fantastic* trades one form of association for a generic positive sentiment, losing the original word's intellectual connotation.

These findings suggest that debiasing is not a simple erasure but a complex transformation that can replace one set of spurious correlations with another. The persistence of some patterns, even after 35 iterations, indicates that certain gender associations may be fundamental to the semantic organization of the corpus. This suggests that post-hoc debiasing may be insufficient to correct for biases deeply embedded in the source data.

Male	Female	Podcasts #1	Debiased #1	Podcasts #2-5	Debiased #2-5
carpentry	sewing	nannying	woodworking	joinery, woodshop, ceramics, metalwork	blacksmithing, carpenter, woodworker, landscaping
surgeon	nurse	surgeons	surgeons	nurse, neurosurgeon, obstetrician, obgyn	neurosurgeon, oculoplastic, surgery, cardiothoracic
burly	blond	mousy	hulking	gray_hair, buxom, matronly, voluptuous	muscly, cis_white, muscular, paunchy
chuckle	giggle	giggle	laugh	laugh, giggling, laughed, giggled	giggle, laughed, laughing, cry_emoji
snappy	sassy	sassy	horseradish	bitchy, snappier, spiced, snippy	queen_elisa, raddish, susie_smith, punchy
football	volleyball	soccer	basketball	lacrosse, volleyball, netball, cheerleading	soccer, rugby, lacrosse, volleyball
physician	registered nurse	obgyn	physicians	nurse, physicians, internist, practitioner	doctor, clinician, holistic_medical_center, internist
architect	interior designer	designer	architects	architects, architecture, architectural, decorator	katherine_nikolai, evan_troxel, engineer, robert_simione
conservatism	feminism	feminism	remaking_one_-nation	progressivism, liberalism, feminist, feminists	conservatives, roy_stewart, progressivism, conservative
guitarist	vocalist	singer	drummer	vocalist, drummer, bassist, songwriter	bassist, singer, vocalist, guitarists
superstar	diva	supermodel	megastar	bianca_belair, grace_davis, ronda_rousey, becky_lynch	superstars, superstardom, derek_green, joe_enlai
pizza	cupcakes	pizzas	pizzas	pepperoni, calzone, pascuali, enchiladas	pepperoni, squalice, mazios, little_caesars
shopkeeper	housewife	lady	kirana	her, housemaid, receptionist, cashier	shopkeepers, tlg, storekeeper, cashier
baseball	softball	softball	basketball	major_league_baseball, mlb, the_major_league_baseball, soccer	mlb, major_league_baseball, football, softball
pharmaceuticals	cosmetics	pharmaceutical	pharmaceutical	medicines, medications, pharma, medication	pharma, medicines, sudical, sun_pharmaceuticals
lanky	petite	gawky	gangly	brunette, mousy, tomboyish, petite	tall, stocky, scrawny, rangey
affable	charming	vivacious	personable	sassy, loquacious, charmingly, demure	likable, charming, sociable, gregarious
brilliant	lovely	fabulous	fantastic	lovely, fantastic, gorgeous, wonderful	phenomenal, superb, amazing, terrific

Table 14. Comparison of gender analogies in original and debiased SPoRC Word2Vec embeddings.