Podcasts as Data: Building a Dataset for Large-Scale Audio Content Analysis

Loren Verreyen[1,2,3]

[1] Antwerp Center for Digital Humanities and Literary Criticism (ACDC), University of Antwerp, Antwerp, Belgium
[2] Amsterdam School for Cultural Analysis (ASCA), University of Amsterdam, Amsterdam, the Netherlands
[3] Research Foundation Flanders (FWO), Brussels, Belgium

Abstract

In recent years, podcasts have rapidly emerged as a popular medium worldwide. In addition to their easy accessibility online, podcasts are characterised by an intimate, aural delivery of their narrative content, creating exciting opportunities for digital scholarship into storytelling. Computational literary studies, however, still remain (hyper)focused on written text, in spite of various calls for more multimodal research. This paper documents the construction of an open-source dataset of textual transcriptions, drawn from a large, representative sample of contemporary, English-language podcasts. The dataset covers a total of 412 days of audio content, transcribed and made available as bag-of-words frequency tables. By making these materials accessible, I hope to stimulate future study of podcasts using techniques from (computational) literary studies. Preliminary experiments suggest that podcast categories in particular offer a fruitful avenue to explore the various textual modes in which podcast producers successfully cater for a diverse global audience.

Keywords: podcasts, audio transcription, genre classification

1 introduction

Over the past two decades, the podcast medium has evolved from a niche audio format into a mainstream multimedial platform.[1] Yet despite its cultural prominence and compelling characteristics, such as serialized storytelling structures and genre-blending possibilities, podcasts remain largely absent from large-scale computational analysis in the humanities. This absence is consistent with a wider methodological bias towards textual data, exemplified by substantial corpora built from novels, news articles, and social media posts that have become central to the Digital Humanities (e.g. [15], [20], [1]). Both technical and institutional barriers contribute to this disparity. On the one hand, podcast's audio based nature fundamentally complicates computational analysis, as employing computational methodologies requires integrating speech and audio processing techniques to handle non-textual data representations. On the other hand, commercial podcast directories such as Apple Podcasts and Spotify implement restrictive policies on bulk data access, complicating the process to assemble representative datasets at scale.

This paper addresses these challenges by introducing a new open-source dataset of transcribed English-language podcasts. I present a dataset of over 15,000 episodes across 1,900 shows, covering all 19 genres supported by Apple Podcasts, representing 412 days of audio. The contribution

Loren Verreyen. "Podcasts as Data: Building a Dataset for Large-Scale Audio Content Analysis ." In: *Computational Humanities Research 2025*, ed. by Taylor Arnold, Margherita Fantoli, and Ruben Ros. Vol. 3. Anthology of Computers and the Humanities. 2025, 212–229. https://doi.org/10.63744/QgeF94c0fP7D.

[1] The dataset introduced in this paper is available at https://doi.org/10.5281/zenodo.17469222

Dataset	Spotify	SPoRC	PodcastRE	Presented dataset
License	CC BY 4.0	Custom License	No formal license	CC BY-SA
Time Period	Jan 2019 - March 2020	May - June 2020	2007 - May 2023	Aug 2019 - April 2025
Size	105K+ episodes 18K+ shows	1.1M+ episodes 247K+ shows	5M+ episodes 1K+ shows	15K+ episodes 1.9K shows
Transcriptions	All episodes	All episodes	± 5000 episodes	All episodes
ASR Model	Google Cloud Video	OpenAI Whisper	[?]	HuggingFace Distil Whisper large-v3

Table 1: Overview Podcast Datasets

of this paper is threefold. First, I describe the methodology for working around data access restrictions imposed by commercial podcast platforms. Second, I provide a comprehensive evaluation of audio transcription quality using automatic speech recognition (ASR) models. Third, I demonstrate possible use cases through preliminary computational analysis, showing how the dataset can be leveraged for scholarly investigation into podcast content across genres.

2 Overview Existing Podcast Datasets

In December 2020, Spotify introduced *100,000 Podcasts: A Spoken English Document Corpus*, the first large-scale dataset of transcribed podcasts [5]. Other notable podcast datasets are SPoRC (Structured Podcast Research Corpus), which includes transcripts of all English-language podcast episodes released between May and June 2020 that are made available through public RSS feeds [9], and PodcastRE (Podcast Research), a dataset and user interface focusing on metadata and audio preservation [12]. A detailed comparison of these datasets alongside the dataset presented in this work is provided in Table 1. As of December 2023, Spotify has discontinued access to their dataset, citing ”shifting priorities” [18]. Also in 2023, PodcastRE posted online that several features (such as researcher access, a specialized user status providing expanded access to the data) would be inconsistently available due to ongoing upgrades to its back-end infrastructure and workflows [11]; a notice that, at the time of writing, remains in effect. As such, access to some of the most prominent podcast datasets has become increasingly limited or inconsistent.

Despite the limitations, these datasets offer(ed) a valuable window into the diverse and dynamic podcasting ecosystem, either through their extended temporal audio coverage (PodcastRE) or the volume of transcripts captured within a short time frame (SPoRC and the 100,000 Podcasts corpus). The dataset presented in this paper builds on these efforts by providing a more comprehensive temporal overview of podcasting practices. Rather than capturing a large number of episodes from a short period, it offers a representative selection of transcripts from a broader time span, allowing for a more nuanced understanding of changing podcast content over time.

3 Metadata Collection

Despite the free accessibility of most podcasts, collecting a representative dataset for scholarly analysis presents challenges. Major directories like Apple and Spotify are commercial platforms that restrict comprehensive data access, which complicates large-scale academic research. The following sections present a methodology for working around these constraints, describing how I systematically identified, accessed, and transcribed podcasts hosted on Apple Podcasts between 2019 and 2025. Podcast makers use podcast hosting platforms such as Spotify for Podcasters, Podbean, and Buzzsprout to store, manage, and distribute their audio files. These podcast hosting platforms create unique RSS feeds for each podcast, capturing essential metadata such as the podcast title, description, category, and author information. RSS, short for Rich Site Summary or Really Simple Syndication, is a standardized XML-based format that allows content to be automatically distributed and updated across multiple platforms. In turn, these RSS feeds are collected

and displayed by commercial podcast directories, such as Spotify, Apple Podcasts (previously a part of iTunes), and YouTube, enabling listeners to stream or download podcasts directly from their preferred platform. Apple's dominance in the podcasting world encompasses both its current industry leadership and its formative influence on the medium. First, alongside Spotify and YouTube, Apple Podcasts remains one of the most prominent platforms for podcast distribution, maintaining a large user base and industry influence [15]. Second, introduced in 2012 as one of the earliest major podcast distribution platforms, Apple Podcasts was instrumental in establishing the technical standards and conventions that continue to shape podcast RSS structure today. This foundational influence is directly reflected in the widespread adoption of RSS feed namespaces such as itunes:title, itunes:category, and itunes:summary, specialized metadata tags that originated with Apple's platform and have become industry standards, representing a level of technical influence that no other platform has achieved to the same degree. Given Apple's persistent and substantial influence in the podcasting landscape from both historical and contemporary perspectives, Apple Podcasts serves as the starting point in the data collection process for this research.

As mentioned before, Apple does not support bulk downloading of podcasts made available through Apple Podcasts. To circumvent this restriction, I developed a pipeline that combines the Podcast Index [5] with Apple's iTunes Search API.[2] The first stage of the pipeline utilizes the Podcast Index, an open and decentralised directory created in 2020 by Adam Curry and Dave Jones that pushes back against Apple's dominant control over the podcasting landscape. An example of this control can be seen in 2018 when Apple removed several podcasts from Apple Podcasts that featured far-right conspiracy theorist Alex Jones (Apple later reversed this ban). Curry and Jones define the directory as "an open, categorised index that will always be available for free, for any use" [5]. The Podcast Index directory contains all podcasts listed on Apple's platform, as well as additional podcasts manually added by users. Its comprehensive scope makes it possible to identify all podcasts available through Apple Podcasts. Podcasts listed on the Podcast Index that are also available on Apple Podcasts are identifiable by their iTunes ID, which acts as a stable identifier.

The second stage of the pipeline queries Apple's iTunes Search API directly using the iTunes IDs collected from the Podcast Index. The Podcast Index is used as an intermediary, but not the final resource. While the directory stores many metadata similar to Apple Podcasts, the Podcast Index applies undocumented preprocessing steps to its metadata, resulting in significant information loss. One of the most problematic preprocessing steps is that the Podcast Index does not allow words to appear twice in the list of genres and, on top of this, splits genre labels. These modifications lead to critical data distortion. For example, science fiction is a subgenre of fiction, but science is also a genre separate from fiction. In the Podcast Index, "science + fiction" could refer to the combination of fiction + science fiction, or fiction + science, or fiction + science fiction + science. Not only is the information lost on what genres were used exactly, but also which genre was selected by the creator as the primary genre (genre listed first). Was this science fiction, or science? This information is important to retain because of the way the Apple Podcasts platform works: podcast makers can choose up to two categories to define their podcast, with the platform supporting nineteen different genres.[3] The category listed first is the primary category, which determines placement in category pages, top charts, and personalized recommendations within Listen Now. Secondary categories influence discoverability through recommendations and search algorithms. If a subcategory is added to the primary category, this subcategory will function as the new primary category. Apple encourages its users to add subcategories to their podcast [2]. Since primary genre classification directly affects podcast visibility and discoverability on the platform, preserving this information accurately is essential for analysis.

[2] https://performance-partners.apple.com/search-api

[3] Arts, Business, Comedy, Education, Fiction, Government, History, Health & Fitness, Kids & Family, Leisure, Music, News, Religion & Spirituality, Science, Society & Culture, Sports, Technology, True Crime, TV & Film

Therefore, I use the iTunes IDs collected from the Podcast Index to crawl data directly from Apple Podcasts using Apple's iTunes Search API. The Podcast Index lists over 4 million podcasts, with more than 2.4 million tagged as English. Among these, over 1.5 million podcasts have an assigned iTunes ID, indicating they are also available on Apple Podcasts. The metadata of all English-tagged podcasts with iTunes ID was collected through the Apple iTunes Search API. For each podcast, detailed information is gathered, including: (1) name podcast, (2) name host, (3) description, (4) number of episodes, (5) date first episode, (6) date most recent episode, (7) primary genre, (8) genres, (9) link to the RSS feed. This approach ensures that the data collection remains consistent with the podcasts listed on Apple Podcasts and leverages the open data accessibility of the Podcast Index to initiate the process while avoiding the data issues caused by its preprocessing steps.

4 Genre-Stratified Subset for Podcast Transcription

This dataset is embedded in a broader PhD project investigating fictionality in true crime podcasts. The relationship between true crime and fiction remains contested within literary and media studies, with scholars proposing different approaches to defining the genre. Worthington offers a definition that collapses the fiction/non-fiction distinction, arguing that "crime fiction refer[s] to all literary material, fiction or fact, that has crime, or the appearance of crime, as its centre and its raison d'être." [21] In contrast, Seltzer maintains the distinction, asserting that "true crime is crime fact that looks like crime fiction" [17], thus acknowledging the genre's reliance on fictional narrative techniques while preserving its factual status. This tension between fictionality and factuality has long characterized true crime, but serialized audio storytelling can intensify it through narrative pacing, suspenseful structure, and intimate aural delivery. My PhD project examines the degree to which true crime podcasts employ narrative and linguistic devices typically associated with fictional storytelling, characterizing the formal and rhetorical strategies that distinguish true crime from other non-fiction genres. The dataset provides the large-scale textual corpus necessary for such comparative analysis across multiple genres.

Yet a caveat is in order: below and elsewhere, the terms "genre" and "category" are often treated as close proxies. This almost interchangeable use of the terms is also reflected in the Apple metadata, where Apple refers to "genre" in its data structures but uses "category" in its online, front-end communication. This blending of terminology can obscure the cultural significance of various genres, reducing them to marketing labels that may not fully capture their semantic depth. Additionally, podcast creators have the freedom to choose their own category/genre, introducing a level of opportunism in labelling that can skew how content is positioned within the podcasting landscape. This classification system will affect how podcasts are organised and how they reach their audience, influencing the metadata associated with them. Analysing the datasets reveals how podcasts are categorised within this framework in which creators select both primary and secondary categories that dictate visibility and discoverability on the platform.

To ensure consistency in the genre-based analysis discussed below, the dataset only includes podcasts that were published from 2019 onward. This starting point was chosen deliberately to avoid possible complications introduced by Apple's major restructuring of its podcast categories in mid-2019, when new top-level categories such as Fiction, History, and True Crime were added, and others were renamed, removed, or reorganized. By beginning the data collection process after these changes were implemented, I maintain a stable and coherent genre classification across the sample. From the full metadata dataset of over 1.5 million English-language podcasts with iTunes IDs, a stratified sample was constructed by selecting 100 podcasts per primary genre. Not all podcast episodes were downloaded for each show. Instead, a second sampling layer was applied at the episode level, randomly selecting 10 episodes for transcription. If a podcast had fewer than 10 episodes, its entire catalog was included, resulting in a genre-balanced corpus. All audio files

Category	Total duration all episodes in hours	Mean duration per episode in minutes	# Episodes	# words
Arts	474	36	795	4,654,890
Business	445	31	858	4,678,750
Comedy	639	47	816	6,612,481
Education	374	27	837	3,739,266
Fiction	370	28	800	3,281,228
Government	476	34	845	4,534,898
Health & Fitness	459	33	827	4,345,757
History	465	33	851	4,327,426
Kids & Family	363	25	861	3,659,109
Leisure	639	47	819	6,343,363
Music	636	47	817	5,430,441
News	530	36	847	5,282,728
Religion & Spirituality	508	37	833	4,596,951
Science	488	35	849	4,692,966
Society & Culture	616	43	856	6,173,801
Sports	642	50	776	6,747,116
Technology	455	35	782	4,484,418
True Crime	546	39	833	5,265,348
TV & Film	763	56	814	7,707,154
Total	9,888 (= 412 days)	38	15,716	96,558,091

Table 2: Dataset composition across 19 podcast genres, showing total audio duration, mean episode length, episode count, and total word count per genre based on generated transcriptions .

were accessed through their RSS feeds collected in the previous step. Each RSS feed contains direct links to the episode audio files, embedded in media content tags. By parsing the RSS feeds, audio URLs were extracted and the files systematically downloaded for all selected episodes. This sampling produced a dataset comprising 1,900 podcasts and 15,716 episodes. A more detailed overview of the number of episodes collected per genre can be found in Table 2.

5 Audio Transcription

Each collected audio file underwent automatic transcription, converting podcast episodes from audio to text. While this transformation inevitably sacrificed certain auditory dimensions such as voice, sound effects, and music, I proceeded with transcription for several reasons. First, focusing on transcriptions opens up an entry point through which the content of podcast shows can be accessed and analyzed, providing a first opportunity to investigate how meaning circulates through this form. Second, the relationship between written and spoken text represents a duality worthy of examination. Podcasts occupy a distinctive position between oral and written communication, retaining characteristics of speech, such as pauses, self-correction and listener engagement, even when informed by written scripts or outlines. This hybrid nature aligns with Ong's concept of secondary orality, which refers to forms of communication shaped by writing yet manifested through oral media such as radio and television. Unlike primary orality, which emerges in cultures, un-

Metric	***Whisper-large-v3***	***Distil-large-v3***
WER (%) ↓	19.19	11.54
ROUGE ↑	91.10	94.60
BLEU ↑	86.85	94.50
∼ time (min.) ↓	40	16

Table 3: Performance Comparison of Whisper Models

touched by writing, secondary orality is "essentially a more deliberate and self-conscious orality, based permanently on the use of writing and print," while maintaining "striking resemblances to the old [primary orality] in its participatory mystique, its fostering of communal sense, its concentration on the present moment, and even its use of formulas" [13]. Podcasts exemplify this distinctive nature, combining narrative involvement with informational delivery in ways that differentiate the medium both from conversation and traditional written text [3]. By transcribing podcasts, converting this textually grounded speech back into written form, the dual nature becomes analytically visible, revealing the interplay of orality and literacy embedded in the medium, offering an opportunity to explore how modern narratives navigate the boundary between spoken and written expression.

To automatically transcribe all podcast episodes, I experimented with OpenAI's automatic speech recognition (ASR) model *Whisper* (*Whisper-large-v3*) [16] and *HuggingFace*'s distilled version of the same model (*Distil-large-v3*) [6]. Both models were selected as they achieved the lowest word error rate (WER) within their respective model families. Since each podcast show generally features a different host, resulting in a wide range of voices across the dataset, the models were not fine-tuned. A small dataset was compiled to evaluate both models, comprising two True Crime podcasts (*Serial*; *Bear Brook*) and two Fiction podcasts (*The Magnus Archives*; *The Black Tapes*). These podcasts were selected due to the availability of publicly accessible transcripts. For each podcast, the audio and transcript of the first five full episodes were collected, resulting in 719 minutes/12 hours of audio, of which 404 minutes were fiction and 315 minutes true crime. The collected ground truth consists of 65,216 words of true crime podcast transcriptions and 47,888 words of fiction podcast transcriptions. Where possible, the creators of the transcripts were contacted to verify their transcription methods, as evaluating ASR models against transcripts potentially generated by similar systems would risk circularity and undermine the evaluation's validity. However, no responses were received. Consequently, I manually verified the accuracy of two episode transcripts per podcast as a quality check. Following the validation of the reference transcripts, both models were evaluated on the dataset, and their transcription performance was assessed using Word Error Rate (WER), ROUGE, and BLEU metrics (Tab. **??**). A lower WER indicates higher transcription accuracy, while higher ROUGE and BLEU scores reflect greater overlap with reference transcripts and, thus, better performance, emphasizing precision and recall, respectively. To my best knowledge, this is the first study on podcast transcripts that evaluates the performance of the selected ASR models prior to applying them at scale. While ASR models have been used in previous podcast studies, existing work typically omits detailed reporting on model accuracy or performance before large-scale deployment.

Hugging Face's *Distil-large-v3* outperforms OpenAI's *Whisper-large-v3* model in both transcription accuracy and efficiency. Specifically, *Distil-large-v3* achieves lower Word Error Rate (11.54 vs. 19.19), and higher ROUGE (94.60 vs. 91.10) and BLEU (94.50 vs. 86.85) scores, indicating superior alignment with the reference transcripts. In addition, the model transcribes the dataset in less than half the time (16 minutes compared to 40 minutes), highlighting a significant advantage in computational efficiency. The comparative advantage of Hugging Face's model is

further underscored when considering the open and community-oriented ethos that characterizes the platform, as opposed to the focus of OpenAI on profit and proprietary control. Genre-specific WER evaluation (True Crime: 12.72, Fiction: 10.19) revealed minimal variation from the overall average of 11.54, confirming that genre did not systematically influence model performance. Based on these results, the *Distil-large-v3* model was used to transcribe the full dataset of 15,716 podcast episodes. The size of the resulting transcripts per genre is also reported in Table 2.

A manual inspection of the transcripts generated by both models revealed that *Whisper-large-v3* falls short when transcribing more challenging audio, for example, recorded phone calls. One of the novelties of the well-known true crime podcast *Serial*, in which Sarah Koenig revisits the 1999 murder of Hae Min Lee by reconstructing the trial and conviction of Lee's ex-boyfriend Adnan Syed, is that Koenig involves Syed in the story by calling him in prison. Many of these phone calls, while important features of the podcast, are absent in the *Whisper-large-v3* transcript. A comparison of both models transcribing the same phone call (S.1, Ep.5, 1:41) can be seen in Table 3. However, the transcription issues observed in *Whisper-large-v3* are not always the result of audio quality alone, nor are they fully systematic. In some cases, the model appears to halt transcription altogether before resuming after a significant delay. This inconsistent behaviour results in large gaps in the output, even when the audio source is relatively clear and continuous. An example of this is shown in Table 4, which compares transcripts of *The Magnus Archives* (S1, Ep.5, 2:34), where part of a coherent and uninterrupted speech segment is missing from the transcript produced by *Whisper-large-v3*, whereas the same passage is captured fully by the smaller *Distil-large-v3* model. This suggests that the larger model also struggles with reliably maintaining output across certain segments, independent of audio clarity or speaker variation.

The nature of the ground truth used in this evaluation also complicates the model evaluation. Podcast audio is not static, as dynamic ad insertion allows advertisements to be updated, added, or deleted after publication without the need to re-edit the underlying audio file. An example of this can be found in Table 5. Second, the published transcript does not *have* to be an exact representation of what is said in the episode. Presenters can deviate from the transcript, resulting in the transcription no longer being an exact representation of what is being said. An example of this can be found in Table 6. These dynamics complicate model evaluation considerably. While a 11.54 word error rate may initially appear problematic, comparisons between models prove more informative than assessing each model against an idealized 100% accuracy benchmark.

6 Genre Analysis

To illustrate the analytical potential of the compiled dataset, I present a preliminary computational examination of genre classification across podcasts. A fundamental question in podcast studies concerns "the different genres that make up podcasting, what characterizes those genres." [10] A question that, despite extensive study of genre characteristics in novels (e.g., [7], [14], [19]), remains largely unexplored in podcast studies. As a demonstration of possible applications for the dataset, this analysis investigates whether computational methods can identify distinctive lexical features associated with different podcast genres. In literary studies, genres are increasingly treated not as fixed categories with essential definitions, but as dynamic social and textual constructs. As Underwood argues, rather than seeking a definitive definition of what makes a genre, genres are better understood not through definitive definitions but through empirical analysis of textual patterns, understanding them as "mutable set[s] of relations between works that are linked in different ways, and resemble each other to different degrees" [19]. This family resemblance approach is particularly well-suited to computational investigation through predictive modeling, where textual similarities and differences can be measured systematically across large corpora. As a demonstration of possible applications for the dataset, this analysis investigates whether computational methods can identify distinctive lexical features associated with different podcast genres.

Whisper-large-v3	Distil-large-v3	Ground truth
Then you have to get out of the school parking lot, but the parking lot is encircled by the school bus loop, so you can't get your car out until the buses fill up and leave, which Adnan wrote took about 10 to 15 minutes. **[???]** That's Adnan elaborating on's letter. **[???]** He wrote that in addition, …	Then you have to get out of the school parking lot, but the parking lot is encircled by the school bus loop, so you can't get your car out until the buses fill up and leave, which Anon wrote took about 10 to 15 minutes. **I'm telling you, I wish like, maybe I'll try to draw a picture of it, but if you could just see how Woodlawn High School left out at 2.15**. That's Anon, elaborating on his letter. **You can't just go to your car and leave. It's going to take a few minutes. So it's just a really tight, really window of time, I mean, for this day taking place, right? And I've always, like, in my heart, man, I've always kind of like, I've been on TV before, like, you know, on these eight line or nightline, where someone tries to reenact the crime. And there's a moment where, like, there's someone like, you know what, this crime could not have been committed according to the set of facts. It's like always this moment, right? It's like I visualize because the route, it's just, I don't know. Oh, hey, we're getting ready to go, right? Sorry. Hey, I got to go. All right. Okay, bye. That happens sometimes. The guards come by and you're just done mid-sentence. Anyway, I can pick up from Adon's letter. He wrote that in addition, …**	Then you have to get out of the school parking lot, but the parking lot is encircled by the school bus loop, so you can't get your car out until the buses fill up and leave. Which, Adnan wrote, "took about ten to fifteen minutes." I wish– maybe I'll try to draw a picture of it, but if you could just see how Woodlawn High School lets out at 2:15. That's Adnan elaborating on his letter. You can't just go to your car and leave. It's going to take a few minutes. So it's just a really tight– window of time for this to have taken place. I've always– in my heart– I've always like– I've seen it on TV before like on Dateline or Nightline where someone tries to reenact the crime. There's a moment where there's someone like "you know what? This crime could not have been committed according to this set of facts." There's always this moment where I visualize the route, it's just– Oh hey, were getting ready to go, right. Sorry. Hey, I gotta go. Alright bye. Okay bye! That happens sometimes. The guards come by and you're just done, mid-sentence. Anyway, I can pick up from Adnan's letter. He wrote that in addition, …

Table 4: Transcription Comparison: Whisper Models vs Ground Truth when transcribing a phone call (*Serial*, S.1, Ep.5, 1:41)

Whisper-large-v3	Distil-large-v3	Ground truth
Statement begins. **[???]** pretty decent. At least it is once you throw in the overtime and the bonuses, and once you've done the rounds you're usually off for the day, so you're working fewer hours than your average office monkey …	Statement begins. **I work as a binman for Waltham Forest Council. It's not a bad job really, as long as you can handle the smell in the early mornings, not to mention that when winter really gets going it can be pretty unpleasant. I've had to chip ice off more than a few bins in my time just to get them open. Still, the pace** pretty decent. At least it is once you throw in the overtime and the bonuses, and once you've done the rounds, you're usually off for the day, so you're working fewer hours than your average office monkey …	Statement begins. I work as a bin man for Waltham Forest Council. It's not a bad job, really, as long as you can handle the smell and the early mornings, not to mention that when winter really gets going it can be pretty unpleasant. I've had to chip ice off more than a few bins in my time, just to get them open. Still, the pay's pretty decent; at least it is once you throw in the overtime and the bonuses, and once you've done the rounds you're usually off for the day, so you're working fewer hours than your average office monkey …

Table 5: Transcription Comparison: Whisper Models vs Ground Truth (*The Magnus Archives*, S.1, Ep.5, 2:35)

Distil-large-v3	Ground truth
to my in laws ethan schreier and janet levine for putting me up in Baltimore so many times in the past year cer was a production of this american live and wbez chicago	to my in laws ethan shire and janet leveen for putting me up in Baltimore so many times in the past year **support for serial comes from mailchimp celebrating creativity chaos and teamwork since 2001 mailchimp send better email** serial is a production of this american live and wbez chicago

Table 6: Transcription Comparison: Generated transcriptions vs ground truth transcribing audio that contains advertisement (*Serial*, S.1 Ep.1, 53:38)

Distil-large-v3	Ground truth
when i opened the bag well the decomposed face was looking right at me it was november 1985	when i opened the bag the decomposed face was looking right at me **I couldn t believe that there was a decomposed body looking me right in the face i can picture it right now i can picture exactly what that face how it looked** it was november 1985

Table 7: Transcription Comparison: generated transcriptions vs ground truth transcribing audio that deviates from script (*Bear Brook*, S.1 Ep.1, 13:18)

Given the broader research agenda of this dataset, investigating fictionality in true crime podcasts, this preliminary analysis focuses particularly on how computational methods can illuminate the tension between True Crime and Fiction categories.

First, all podcast transcripts are processed into TF-IDF representations, retaining all words that occur in at least two different podcast shows to filter out show-specific vocabulary. Next, dimensionality reduction through PCA and t-SNE was applied to the training set, resulting in two-dimensional projections that provide a preliminary overview of genre organization and an initial assessment of intra-genre cohesion and inter-genre separation. The t-SNE visualization reveals preliminary indications of genre separability, with Religion & Spirituality (gray) and Sports (orange) standing out in particular, forming relatively distinct clusters (Fig. 8). Next, a linear Support Vector Machine (LinearSVC) classifier was trained on the training set to predict genre labels based on the TF-IDF feature vectors. LinearSVC was selected because it returns interpretable feature coefficients, providing insight into which lexical features are most predictive of each genre, enabling deeper qualitative insight into genre characteristics. The model was evaluated on a held-out test set compiled through an 80/20 train-test split stratified by genre, with entire podcast shows assigned to either the training or test set to prevent data leakage.

When trained on all 19 genres simultaneously, the classifier achieves a macro-average F1-score of 0.48. More detailed metrics for all genres are provided in Table 7. Performance varies substantially across genres. Genres such as Music (F1: 0.71), Sports (F1: 0.73), and Religion & Spirituality (F1: 0.73) display strong lexical coherence, while News (F1: 0.34), Arts (F1: 0.29), and Society & Culture (F1: 0.22) achieve much lower F1-scores. The confusion matrix reveals systematic misclassification patterns for these lower-performing genres: News is frequently confused with Government or Sports; Arts with TV & Film or History; Education is distributed across multiple categories. This pattern suggests that these genres are lexically more diffuse or contain overlapping vocabulary with adjacent genres, making them difficult to distinguish based on lexical information alone. However, the consistent clustering of misclassifications suggests that the classification challenges result from semantic overlap between genres rather than systematic model limitations. The 20 most predictive lexical features for each of the 19 genres are presented in Appendix A.

Given the broader research context of this dataset, I examined True Crime (F1: 0.57) and Fiction (F1: 0.58) more closely. The confusion matrix shows that when True Crime episodes are misclassified, they are frequently misclassified as Fiction (32 out of 60 misclassifications). At the same time, Fiction episodes are most often misclassified as True Crime (28 out of 61 misclassifications). This overlap indicates substantial shared vocabulary between the genres. Taking a closer look at their vocabulary, True Crime is characterized by words like ”crime,” ”prison,” ”murder,” ”evidence,” and ”investigators,” while Fiction features words like ”armor”, ”paranormal”, and ”nightmares”, as well as mutliple proper nouns. Yet the high rate of bidirectional confusion suggests that the fiction/non-fiction distinction does not manifest through abandonment of shared narrative vocabulary, but rather through strategic deployment of specific lexical markers, particularly those signaling reality claims and institutional authority. Both genres may draw on similar storytelling techniques, but True Crime seems to leverage crime-specific vocabulary tied to real institutions and investigative processes to establish factual grounding, while Fiction employs more imaginative worldbuilding terminology. The genres are thus distinguished not by fundamentally different content, but by how they lexically construct and authenticate their narratives. Of course, lexical features alone cannot fully explain the fuzzy boundaries between these genres, neither can they settle the deeper questions about the relationship between true crime narratives and fictionality. This preliminary analysis merely suggests promising avenues for further investigation such as examining narrative structure and rhetorical strategies. The dataset’s strength lies not in resolving these questions but in enabling researchers to pursue them at scale.

In sum, this genre classification analysis offers an initial glimpse into the broader analytical potential of the podcast dataset. While the current approach only provides a first perspective on lexical genre markers, it also opens up several avenues for future research. For instance, given that podcast platforms encourage genre hybridity to cater to highly specific audiences, further work could explore how this hybridity complicates or reshapes conventional genre boundaries. Moreover, understanding genre as a transmedial constructs invites comparative studies across media. For instance, previous research has looked into the lexical features of science fiction in the context of novels [19]. The presence of this genre within the podcasting ecosystem enables cross-media comparisons, making it possible to investigate how genre-specific features persist, transform, or diverge across different media formats, contributing to a richer understanding of genre as a transmedial and dynamic phenomenon.

	precision	**recall**	**f1-score**	**support**
Arts	0.31	0.28	0.29	156
Business	0.52	0.61	0.56	174
Comedy	0.37	0.61	0.46	161
Education	0.34	0.20	0.25	173
Fiction	0.56	0.60	0.58	152
Government	0.42	0.43	0.43	161
Health & Fitness	0.41	0.48	0.44	164
History	0.42	0.33	0.37	181
Kids & Family	0.57	0.56	0.57	185
Leisure	0.47	0.26	0.33	162
Music	0.62	0.82	0.71	163
News	0.33	0.35	0.34	154
Religion & Spirituality	0.74	0.72	0.73	170
Science	0.49	0.48	0.48	172
Society & Culture	0.31	0.17	0.22	162
Sports	0.63	0.85	0.73	151
Technology	0.62	0.61	0.61	165
True Crime	0.52	0.64	0.57	168
TV & Film	0.53	0.71	0.61	167
accuracy			0.50	3141
macro avg	0.48	0.50	0.48	3141
weighted avg	0.48	0.50	0.48	3141

Table 8: Evaluation Genre Classification

7 Discussion

The presented dataset is limited in a number of ways. First, the dataset is limited to English-language podcasts. The dataset presents podcast content in textual form, meaning that audio-specific information such as speaker, voice, musical elements, and sound design is omitted from possible analysis. In future iterations of this dataset, different ways to capture audio-specific information could be explored and applied to enrich the dataset. The SPoRC dataset sets an example here, as the dataset not only shares the text transcripts but also includes audio-specific information such as pitch and speaker diarization [9]. More, the dataset's current iteration starts in 2019 to ensure consistency in genre labeling, as Apple Podcasts introduced a big shift in its supported genres in the summer of that year. The audio collection and transcription processes will be repeated on a

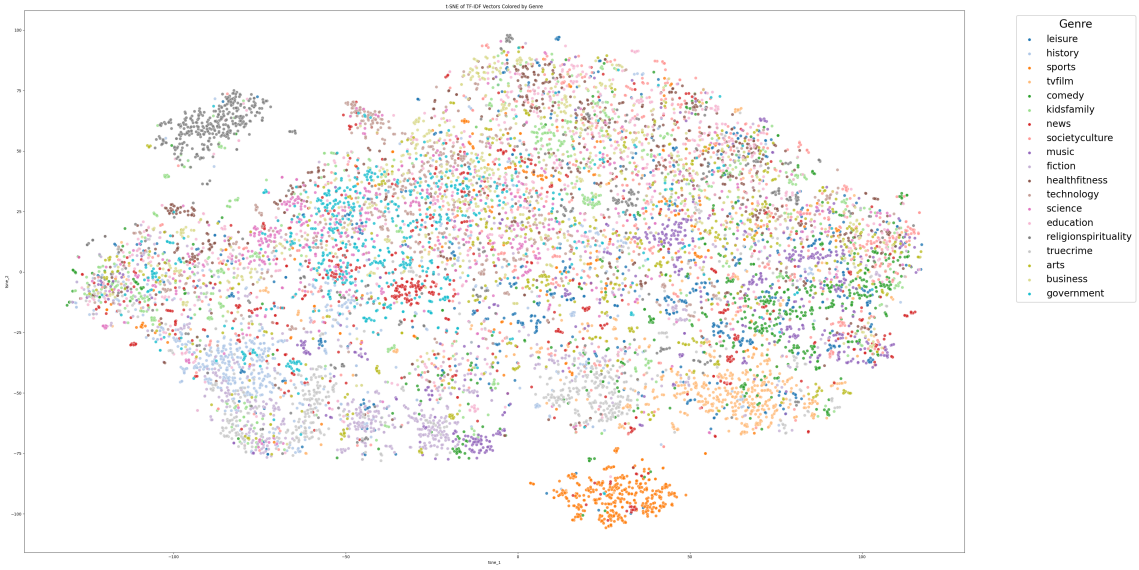

Figure 1: t-SNE genres

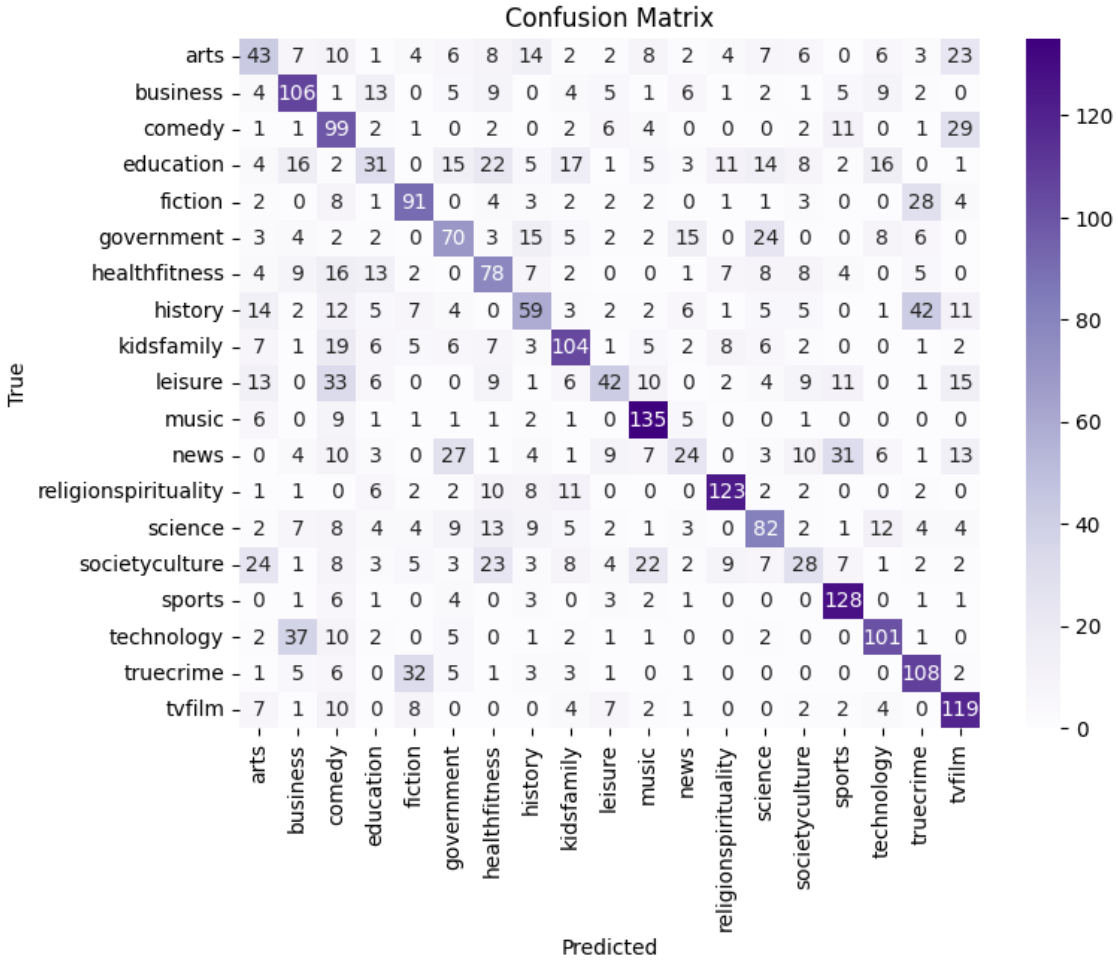

Figure 2: Confusion matrix SVM genre classification

random selection of podcasts published before 2019 to complement the presented dataset. Finally, one question remains: do all collected audio files qualify as podcasts? Podcasting is often celebrated as an accessible and democratized medium, where minimal technical barriers allow a wide range of creators to produce and distribute content [10]. This openness is clearly reflected in the dataset, which includes an array of audio materials, ranging from brief recordings of poetry readings and song covers to hour-long investigative episodes on true crime. Such diversity prompts critical reflection on the nature of the dataset and its alignment with the concept of podcasting. Should episodes that primarily consist of music or other non-spoken content be considered podcasts in the traditional sense, or might their inclusion dilute the focus of the dataset? Spotify, for example, excluded podcasts that contained less than 50% speech when constructing the 100,000 Podcasts corpus [5]. Content-based restrictions, such as excluding files with a majority of musical content, could refine the dataset but risk excluding materials that contribute to the medium's evolving boundaries. The definition of a podcast often centers on its technological characteristics, leaving content characteristics undefined. Bottomley, for example, states that "[t]echnonologically speaking, podcasting refers to digital audio files (e.g. MP3s) delivered via RSS to an Internet-connected computer or portable media player" [4]. Since all audio files included in this dataset were collected through their RSS feeds, they all meet this technological criterion, underscoring the inherent ambiguity in defining what precisely constitutes a podcast episode. This ambiguity invites further consideration of whether technological parameters, content characteristics, or both should guide dataset inclusion criteria.

8 Conclusion

This paper has introduced a large-scale, open-source dataset of podcast transcripts comprising over 15,000 episodes across 1,900 shows, spanning all 19 podcast genres recognized by Apple Podcasts. By converting 412 days of audio content into textual data and distributing it as bag-of-words frequency tables, the dataset makes podcast content accessible for scalable textual analysis while respecting copyright constraints. In doing so, it addresses a significant gap in the Digital Humanities, where computational research has remained overwhelmingly text-centric despite growing scholarly interest in multimodal storytelling.

Throughout this work, I have examined existing podcast datasets and identified several key challenges inherent in collecting podcast data at scale. Using the Podcast Index as an intermediary source, I compiled a comprehensive dataset containing metadata for all English-language podcasts available on Apple Podcasts, before selecting a stratified subset for transcription. The comparative evaluation of transcription models revealed that Hugging Face's lighter *Distil-large-v3* outperformed its larger, proprietary alternative in both accuracy and efficiency. The preliminary genre classification experiment further demonstrates the dataset's potential for exploring linguistic patterns and genre hybridity in podcasting, revealing distinct lexical signatures for certain genres while highlighting the complexity of genre boundaries in this medium.

By making this dataset publicly available, I aim to encourage new lines of inquiry into the cultural, linguistic, and formal dimensions of podcasts. Future work may expand upon this resource by incorporating audio-specific features, extending temporal coverage, and enabling comparative studies across media. Ultimately, I hope this project contributes to the ongoing multimodal turn in the Digital Humanities and helps establish podcasts as a vital and analyzable form of contemporary cultural expression.

Copyright. Although most podcasts can be freely downloaded online, they are protected by international copyright and, under most legal systems, they cannot be freely be redistributed. In order to comply with international copyright law, I have decided to share the resulting transcriptions as bag-of-words frequency table: because this "Derived Data Format" (DDF) does not allow to reconstruct the original transcription [8], this does not violate the copyright of the original podcasts,

but still enables researchers to carry out valid lexical analysis. Importantly, many contemporary corpora are available in a similar format (such as the CONLIT dataset, [15]), so that the aural modality of podcasts, for example, could be compared to more formally published books in print. I share the data under liberal Creative Commons license that should stimulate their re-use in the scientific community.

References

[1] Allés-Torrent, Susanna, Rio Riande, Gimena del, Bonnell, Jerry, Song, Dieyun, and Hernández, Nidia. "Digital narratives of covid-19: A twitter dataset for text analysis in spanish". In: *Journal of Open Humanities Data 7*, no. 0 (2021), p. 5. DOI: 10.5334/johd.28.

[2] Apple Podcasts for Creators. "Apple Podcasts categories". https://podcasters.apple.com/support/1691-apple-podcasts-categories. Accessed: 2025-07-17. n.d.

[3] Babayode, Aminat, Bosman, Laurens, Chan, Nicole, Ehret, Katharina, Fong, Ivan, Harris, Noelle, Hewton, Alissa, Reid, Danica, Taboada, Maite, and Wong, Rebekah. "Structural linguistic characteristics of podcasts as an emerging register of computer-mediated communication". In: *Proceedings of the 10th International Conference on CMC and Social Media Corpora for the Humanities*. 2023, pp. 3–6. DOI: 10.1075/rs.23010.ehr.

[4] Bottomley, Andrew J. "Podcasting: A decade in the life of a 'new' audio medium: Introduction". In: *Journal of Radio & Audio Media 22*, no. 2 (2015), pp. 164–169.

[5] Clifton, Ann et al. "100,000 Podcasts: A Spoken English Document Corpus". In: *Proceedings of the 28th International Conference on Computational Linguistics*, ed. by Donia Scott, Nuria Bel, and Chengqing Zong. Barcelona, Spain (Online): International Committee on Computational Linguistics, Dec. 2020, pp. 5903–5917. DOI: 10.18653/v1/2020.coling-main.519.

[6] Gandhi, Sanchit, Von Platen, Patrick, and Rush, Alexander M. "Distil-whisper: Robust knowledge distillation via large-scale pseudo labelling". https://arxiv.org/abs/2311.00430. 2023. DOI: 10.48550/arXiv.2311.00430.

[7] Hettinger, Lena, Becker, Martin, Reger, Isabella, Jannidis, Fotis, and Hotho, Andreas. "Genre classification on German novels". In: *2015 26th International Workshop on Database and Expert Systems Applications (DEXA)*. IEEE. 2015, pp. 249–253. DOI: 10.1109/DEXA.2015.62.

[8] Kugler, Kai, Münker, Simon, Höhmann, Johannes, and Rettinger, Achim. "InvBERT: Reconstructing Text from Contextualized Word Embeddings by inverting the BERT pipeline". In: *Journal of Computational Literary Studies 2*, no. 1 (2023), pp. 1–18. DOI: 10.48694/jcls.3572.

[9] Litterer, Benjamin Roger, Jurgens, David, and Card, Dallas. "Mapping the Podcast Ecosystem with the Structured Podcast Research Corpus". In: *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, ed. by Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar. Vienna, Austria: Association for Computational Linguistics, July 2025, pp. 25132–25154. ISBN: 979-8-89176-251-0. DOI: 10.18653/v1/2025.acl-long.1222.

[10] McGregor, Hannah. "Podcast studies". In: *Oxford research encyclopedia of literature*. Oxford University Press, 2022. DOI: 10.1093/acrefore/9780190201098.013.1338.

[11] Morris, Jeremy. "About PodcastRE". https://podcastre.org/about. 2014. (Visited on 07/17/2025).

[12] Morris, Jeremy Wade, Hansen, Samuel, and Hoyt, Eric. “The PodcastRE Project: Curating and Preserving Podcasts (and Their Data)”. In: *Journal of Radio & Audio Media 26*, no. 1 (2019), pp. 8–20. DOI: 10.1080/19376529.2019.1559550.

[13] Ong, Walter J. *Orality and Literacy: The Technologizing of the Word*. 2nd. New York, NY, USA: Routledge, 2002. DOI: 10.4324/9780203328064.

[14] Piper, Andrew. “Fictionality”. In: *Journal of Cultural Analytics 2*, no. 2 (2016). DOI: 10.22148/16.011..

[15] Piper, Andrew. “The CONLIT Dataset of Contemporary Literature”. In: *Journal of Open Humanities Data 8*, no. 24 (2022), pp. 1–7. DOI: 10.5334/johd.88.

[16] Radford, Alec, Kim, Jong Wook, Xu, Tao, Brockman, Greg, McLeavey, Christine, and Sutskever, Ilya. “Robust Speech Recognition via Large-Scale Weak Supervision”. https://arxiv.org/abs/2212.04356. 2022. DOI: 10.48550/ARXIV.2212.04356.

[17] Seltzer, Mark. *True crime: Observations on violence and modernity*. New York, NY, USA: Routledge, 2013. DOI: 10.4324/9780203944202.

[18] Spotify. “Spotify Podcasts Dataset”. https://web.archive.org/web/20240729065449/http://podcastsdataset.byspotify.com/. Archived on July 29, 2024. Published by Spotify. Accessed on July 14, 2025. 2024.

[19] Underwood, Ted. “The Life Cycles of Genres”. In: *Journal of Cultural Analytics 2*, no. 2 (2016). DOI: 10.22148/16.005.

[20] Westerling, Kalle, Beelen, Kaspar, Hobson, Tim, McDonough, Katherine, Pedrazzini, Nilo, Wilson, Daniel CS, and Ahnert, Ruth. “LwMDB: Open Metadata for Digitised Historical Newspapers from British Library Collections”. In: *Journal of Open Humanities Data 11*, no. 1 (2025), p. 32.

[21] Worthington, Heather. *Key concepts in crime fiction*. Basingstoke, UK: Palgrave Macmillan, 2011.

A Top Lexical Features by Genre

This appendix lists the top 20 most predictive lexical features (ranked by SVM coefficient) for each of the 19 podcast genres in the dataset.

A.1 Arts

art (3.10), wine (2.82), chef (2.75), poem (2.64), mccarthy (2.45), book (2.34), kombucha (2.32), read (2.25), creative (2.21), broadway (2.16), cooking (2.13), waves (2.05), pj (1.98), manga (1.94), coffee (1.88), show (1.84), style (1.79), reading (1.77), artist (1.74), kroger (1.72)

A.2 Business

marketing (3.30), financial (3.02), business (2.78), career (2.77), quant (2.49), corbell (2.36), emc (2.30), product (1.98), leader (1.90), growth (1.89), dental (1.83), leadership (1.78), corporate (1.73), mindset (1.67), teachers (1.65), returners (1.65), stock (1.63), egan (1.56), hats (1.52), trial (1.45)

A.3 Comedy

comedy (3.38), elena (1.89), english (1.89), podcast (1.74), mr (1.70), improv (1.65), joke (1.58), luke (1.53), zach (1.45), marty (1.43), blunts (1.40), wednesdays (1.38), tell (1.34), ethan (1.33), drinking (1.33), blair (1.31), gonnae (1.31), std (1.30), goes (1.29), s (1.28)

A.4 Education

greenbush (2.62), rally (2.23), norton (2.04), guyana (1.92), excited (1.89), hawking (1.85), royston (1.83), janelle (1.79), yes (1.78), asc (1.76), lateral (1.76), thriving (1.75), conversation (1.72), shirel (1.68), tree (1.67), registrar (1.64), portuguese (1.62), preparedness (1.58), shelley (1.57), u (1.54)

A.5 Fiction

armor (1.76), dr (1.71), nightmares (1.66), asimov (1.62), d (1.60), date (1.60), fiction (1.59), gotcha (1.52), lily (1.51), paranormal (1.49), stan (1.48), leap (1.48), jim (1.48), sauron (1.47), theatre (1.47), serena (1.46), asoka (1.45), wars (1.43), gondor (1.43), season (1.42)

A.6 Government

westchester (2.43), idaho (2.29), court (2.09), coleraine (2.03), mandan (2.03), injury (1.94), community (1.75), government (1.71), state (1.65), summerland (1.64), policy (1.60), fbi (1.60), airmen (1.59), maine (1.56), delaware (1.55), infantry (1.55), party (1.53), chandler (1.53), emergency (1.48), trade (1.48)

A.7 Health & Fitness

wellness (2.79), or (2.31), sunflower (2.28), massage (2.12), keto (2.05), patient (1.88), physician (1.87), dentistry (1.86), more (1.68), radiology (1.58), dementia (1.56), cancer (1.56), na (1.51), breast (1.49), triathlon (1.47), psychology (1.41), bipolar (1.40), medication (1.38), swimming (1.35), denise (1.35)

A.8 History

history (4.25), vintage (2.34), historical (1.95), florida (1.91), viva (1.83), war (1.74), men (1.69), ellsworth (1.68), british (1.64), banting (1.54), ancient (1.54), civil (1.54), iowa (1.48), dance (1.45), guinness (1.44), shara (1.42), diddy (1.40), victorian (1.40), remember (1.40), archaeology (1.40)

A.9 Kids & Family

parenting (4.06), dads (3.18), dog (3.08), kids (3.01), dad (2.56), moms (2.47), homeschool (2.38), fabled (2.21), motherhood (2.03), mommy (2.03), fatherhood (1.96), owl (1.93), mom (1.92), parent (1.88), baby (1.83), avery (1.75), liam (1.61), jp (1.52), shiba (1.51), medical (1.51)

A.10 Leisure

game (2.76), anime (2.45), deck (2.28), ev (2.24), uvm (2.24), schmidt (2.01), springs (1.92), whenever (1.92), devin (1.87), flowers (1.83), fish (1.81), stream (1.78), darren (1.76), creature (1.76), gar (1.73), damage (1.66), motorcycle (1.61), jules (1.60), birds (1.58), birding (1.57)

A.11 Music

guitar (3.12), music (3.03), song (2.89), band (2.74), piano (2.52), bands (2.37), festival (2.21), eurovision (2.11), track (1.95), be (1.91), songs (1.88), playing (1.82), playlist (1.81), stage (1.75), album (1.69), record (1.68), pit (1.64), smoke (1.63), artists (1.62), drums (1.52)

A.12 News

cape (2.35), trump (2.24), politics (2.09), china (2.02), news (2.00), biden (1.94), bonsai (1.82), county (1.80), border (1.77), smith (1.77), on (1.72), newfoundland (1.59), kylie (1.59), labor (1.58), minneapolis (1.57), saying (1.53), conservative (1.53), npr (1.51), custody (1.48), tv (1.47)

A.13 Religion & Spirituality

bible (2.77), kessner (2.76), god (2.57), jesus (2.44), augustine (2.37), lord (2.08), scripture (1.94), says (1.90), hashem (1.66), israel (1.63), when (1.62), church (1.62), christ (1.58), joseph (1.56), mormon (1.54), verse (1.54), goddess (1.54), torah (1.53), scriptures (1.50), sermon (1.49)

A.14 Science

rhino (2.52), environmental (2.47), conservation (2.29), haiku (2.28), science (2.25), transportation (1.97), beatrice (1.93), lenses (1.92), mortuary (1.80), brain (1.78), scientists (1.74), cma (1.73), pharmacy (1.72), ocean (1.69), nurses (1.66), lab (1.59), disease (1.57), transit (1.56), oceans (1.55), biotech (1.54)

A.15 Society & Culture

pittsburgh (2.65), whiskey (2.58), diaries (2.28), cuba (2.28), abortion (2.25), travel (2.22), cincinnati (1.80), aesthetic (1.77), kd (1.74), friendships (1.66), things (1.64), cave (1.63), salon (1.60), utah (1.58), yo (1.58), girlfriends (1.49), renaissance (1.47), y (1.47), racism (1.44), bryce (1.43)

A.16 Sports

race (2.57), climbing (2.48), coach (2.34), horses (2.00), soccer (1.99), surf (1.95), spartan (1.92), players (1.91), league (1.88), skate (1.83), horse (1.80), season (1.79), surfing (1.79), team (1.69), teams (1.69), run (1.68), boards (1.62), draft (1.60), running (1.55), coaches (1.48)

A.17 Technology

tech (3.57), blockchain (2.23), tesla (2.04), vr (1.97), users (1.95), ai (1.88), mac (1.82), code (1.80), haptics (1.78), technology (1.77), linkedin (1.75), webflow (1.67), freight (1.67), web (1.66), strife (1.65), impromptu (1.64), user (1.62), actually (1.61), software (1.60), engineering (1.58)

A.18 True Crime

crime (4.51), prison (2.99), murder (2.88), case (2.56), shadow (2.28), police (2.14), caviar (1.91), fire (1.86), found (1.75), deer (1.73), missing (1.62), investigators (1.60), spotlight (1.58), investigation (1.53), num (1.43), jimmy (1.41), dna (1.40), quote (1.35), investigator (1.34), story (1.34)

A.19 TV & Film

film (3.83), movie (3.54), scene (3.03), trek (2.60), barney (2.53), horror (2.41), episode (2.40), hollywood (2.27), um (1.95), character (1.83), actor (1.66), films (1.63), show (1.53), issa (1.51), survivor (1.49), production (1.49), anyways (1.49), odo (1.45), minute (1.43), walt (1.41)