Masculine Defaults via Gendered Discourse in Podcasts and Large Language Models

Maria Teleki, Xiangjue Dong, Haoran Liu, James Caverlee

Texas A&M University
{mariateleki, xj.dong, liuhr99, caverlee}@tamu.edu

Abstract

Masculine defaults are widely recognized as a significant type of gender bias, but they are often *unseen* as they are *under-researched* (Cheryan and Markus 2020). Masculine defaults involve three key parts: (i) the cultural context, (ii) the masculine characteristics or behaviors, and (iii) the reward for, or simply acceptance of, those masculine characteristics or behaviors. In this work, we study *discourse-based masculine defaults*, and propose a twofold framework for (i) the large-scale discovery and analysis of gendered discourse words in spoken content via our *Gendered Discourse Correlation Framework (GDCF)*; and (ii) the measurement of the gender bias associated with these gendered discourse words in LLMs via our *Discourse Word-Embedding Association Test (D-WEAT)*. We focus our study on podcasts, a popular and growing form of social media, analyzing 15,117 podcast episodes. We analyze correlations between *gender* and *discourse words* – discovered via LDA and BERTopic – to automatically form gendered discourse word lists. We then study the prevalence of these gendered discourse words in domain-specific contexts, and find that gendered discourse-based masculine defaults exist in the domains of business, technology/politics, and video games. Next, we study the representation of these gendered discourse words from a state-of-the-art LLM embedding model from OpenAI, and find that the masculine discourse words have a more stable and robust representation than the feminine discourse words, which may result in better system performance on downstream tasks for men. Hence, men are rewarded for their discourse patterns with better system performance by one of the state-of-the-art language models – and this embedding disparity constitutes a representational harm and a masculine default.

Introduction

Masculine defaults are a type of gender bias "in which characteristics and behaviors associated with the male gender role are valued, rewarded, or regarded as standard, normal, neutral, or necessary aspects of a given cultural context" (Cheryan and Markus 2020). Hence, there are three parts to a masculine default: (i) the cultural context, (ii) the male characteristics or behaviors, and (iii) the reward for, or simply acceptance (neutral) of the male characteristics or behaviors. **Hence, to determine whether or not a behavior constitutes a masculine default (Cheryan and Markus 2020), we consider: What is the reward or standard associated with a given masculine behavior?** For example, in the cultural context of the United States, the prevalence of men in computer science is a masculine default, as men are economically rewarded for being computer scientists via statistically higher salaries (U.S. Bureau of Labor Statistics 2023), and are largely socially accepted in this role (Cheryan and Markus 2020). This masculine default, then, propagates social injustice, as "women feel a lower sense of belonging and anticipate less success" in computer science, and do not enter the field at a comparable rate to men and reap the economic rewards (Cheryan and Markus 2020; American Society for Engineering Education 2022). These masculine defaults result in the *other-ing* of women (Beauvoir 1949).

Considerable prior research has examined gender differences in social media (e.g., Wang, Pappu, and Cramer (2021); Kalhor et al. (2023); Johnson et al. (2021); Wang and Horvát (2019)) and in LLMs (e.g., Dong et al. (2023); Caliskan, Bryson, and Narayanan (2017); May et al. (2019); Bolukbasi et al. (2016)). But how do *masculine defaults* manifest on social media? And how do they impact emerging systems like large language models (LLMs) that are trained in part over social media? While masculine defaults are highly connected to gender differences,[1] there is a research gap in identifying and analyzing masculine defaults that arise through *gender differences in discourse*.

Specifically, we focus on patterns of discourse in spoken communication, including fillers (e.g., *uh*, *um*), discourse markers (e.g., *well*, *you know*, *I mean*), false starts (e.g., *It was, anyways, I went to Target yesterday*) and more (Merriam-Webster 2024; Shriberg 1994). Such discourse

[1] We consider the binary definitions of sex (female/male) and gender (women/men, feminine/masculine) in our work due to (i) continuity with previous work in the gender debiasing task in the NLP community (Caliskan, Bryson, and Narayanan 2017; Bolukbasi et al. 2016), and (ii) modeling constraints – i.e., *inaSpeechSegmenter* (Doukhan et al. 2018a) for gender approximation via audio signal. This definition, however, is not representative of the sex and gender spectrums – and transgender, intersex, intersectional identities, and other identities are also not represented in this binary definition (Ghai, Hoque, and Mueller 2021; Ovalle et al. 2023; Seaborn, Chandra, and Fabre 2023). This is an important direction for future work.

words are non-content related words that serve important social purposes with respect to gender, such as to *"hold the floor"* in conversation (Shriberg 1994, 1996). Previous work notes gender differences in how men and women use specific types of *discourse words* – for example, men use more filled pauses and repeats (Shriberg 1996; Bortfeld et al. 2001) than women. However, these studies lack an automated method for large-scale discourse word discovery and gender analysis, primarily relying on the Switchboard corpus (Mitchell et al. 1999) – an older, human-annotated corpus which is not representative of the range of natural speech patterns, as the phone calls were recorded in the manufactured, awkward situation of randomly-pairing two callers and assigning them a topic to discuss.

Hence, we propose in this paper a twofold framework for (i) the large-scale discovery and analysis of gendered discourse words in spoken content via our **Gendered Discourse Correlation Framework (GDCF)**; and (ii) the measurement of the gender bias associated with these gendered discourse words in LLMs via our **Discourse Word-Embedding Association Test (D-WEAT)**. Concretely, we focus our study on podcasts, a popular and growing form of social media (Clifton et al. 2020). According to Pew Research, "42% of Americans ages 12 and older have listened to a podcast in the past month" as of 2023 compared to 12% in 2013 (The Pew Research Center 2023). We analyze the *rewards* associated with *gendered discourse words* in 15,117 podcast episodes from the Spotify Podcast Dataset (Clifton et al. 2020) – i.e., discourse words with significant positive correlations with either men or women – to determine whether or not masculine defaults are present. Our study is organized around the following research questions:

- *RQ0: How are women and men's discourse different?*
- *RQ1: Are discourse-based masculine defaults present in domain-specific contexts?*
- *RQ2: Are discourse-based masculine defaults present in LLM embeddings?*

We first (RQ0) introduce our *Gendered Discourse Correlation Framework (GDCF)*, a framework for discovering gendered discourse words, with features which are centered around spoken content – specifically, an audio-based GENDER SEGMENTER (Doukhan et al. 2018a), a TOPIC MODELER via LDA (Blei, Ng, and Jordan 2003) and BERTopic (Grootendorst 2022), and a specialized CONVERSATIONAL PARSER (Jamshid Lou and Johnson 2020). We analyze correlations between *gender* and *discourse words* to automatically form gendered discourse word lists. Additionally, GDCF is a flexible framework which can be extended to other forms of audio speech data – such as short videos that are prevalent on TikTok, Instagram, and YouTube, long videos on YouTube, streamers on Twitch, and more.

We then study (RQ1) the prevalence of these gendered discourse words in domain-specific contexts. We find that masculine discourse words are positively correlated with the business domain. Because participation in the business domain grants economic rewards, there are indeed discourse-based masculine defaults present in the business domain. We additionally show that this is the case for the domains of technology/politics and video games, and provide more, related results in the Appendix.

Next, we study (RQ2) the representation of these gendered discourse words in a state-of-the-art LLM embeddings model from OpenAI, `text-embedding-3-large`. We find that the masculine discourse words have a more stable and robust representation than the feminine discourse words, resulting in better system performance on downstream tasks for men. Hence, men are rewarded for their discourse patterns with better system performance by one of the state-of-the-art language models – and therefore this difference in the embedding representations for women and men constitutes a *representational harm* (Blodgett et al. 2020) and a masculine default.

We release our code at https://github.com/mariateleki/masculine-defaults and the extended results at https://www.gendered-discourse.net.

Related Work

Sex, Gender, and Language. We focus in our work on *gender* rather than *sex*:[1] sex (female/male) is established based on biology; whereas, gender (women/men, feminine/-masculine) "is the activity of managing situated conduct in light of normative conceptions of attitudes and activities appropriate for one's sex category" (West and Zimmerman 1987; Unger 1979; Muehlenhard and Peterson 2011). Gender is something that people "do," and "gender [can be understood] as a routine, methodological, and recurring accomplishment" (West and Zimmerman 1987). In Butler's theory of *gender performativity*, "[g]ender is instituted through the stylization of the body and, hence, must be understood as the mundane way in which bodily gestures, movements, and enactments of various kinds constitute the illusion of an abiding gendered self" (Butler 1988) – an *enactment*, then, includes *language*: the way women and men speak. Butler argues that conforming to this *gender schema* – wherein certain "attitudes," "activities," "attributes," and "behaviors", including language, are assigned to either women or men (Bem 1984; West and Zimmerman 1987) – is necessary for women to "ask for recognition in the law or in political life" (Butler 2009). In this way, masculinities and femininities relate to social and political power.

Hegemonic masculinity refers to a performative, "'currently accepted' strategy" for maintaining the patriarchal imbalance of social and political power via cultural dominance (Connell 1995, 1987). Maintaining power necessitates different strategies over time, thus, hegemonic masculinity is highly contextual. One recent type of hegemonic masculinity is *technomasculinity* – the form of masculinity associated with high-tech professions, such as engineering and science (Cooper 2000; Lockhart 2015; Bulut 2020; Goree, Crandall, and Su 2023). Hence, as gender and gender roles are highly contextual, we limit our definition of gender temporally, to recently, and geographically, to the United States.

Hegemonic masculinity, then, is closely related to *masculine defaults*, which are a form of *other-ing* – conciously and/or subconciously – that occurs as the result of the masculine social and political hierarchy. "Masculine defaults include ideas, values, policies, practices, interaction styles,

norms, artifacts, and beliefs that often do not appear to discriminate by gender but result in disadvantaging more women than men" (Cheryan and Markus 2020). Masculine defaults relate to other-ing in that "alterity is the fundamental category of human thought" and "He is the Subject; he is the Absolute. She is the Other" (Beauvoir 1949). In other words, he is the *default*, and she is the other. An example of a masculine default in language is the use of *masculine generics*, as "[a]n almost universal and fundamental asymmetry lies in the use of masculine generics. In English, for example, generic he can be used when gender is irrelevant (e.g., the user... he)" rather than 'she' (Sczesny, Formanowicz, and Moser 2016). For a masculine behavior to be considered a *masculine default*, there must be a *reward* or a *standard* associated with the use of the masculine behavior (Cheryan and Markus 2020).

Podcast Language Analysis. Podcasts have come under increased research scrutiny in the past few years. For example, Yang et al. (2019) analyzed non-textual characteristics of podcasts (like energy or seriousness) through audio spectrogram representation learning methods. Clifton et al. (2020) conducted an analysis of the Spotify dataset podcasts, where they also found that discourse topics exist, and they found a higher frequency of first-person pronouns and amplifiers as compared to the Brown corpus. Valero, Baranes, and Epure (2022) studied topic modeling on podcasts for information retrieval with the Spotify dataset. Martikainen, Karlgren, and Truong (2022) have examined how stylistic features relate to genres on a small scale using PCA and k-means clustering: they analyzed a subset of 14 episodes then a subset of 911 episodes. They also used *inaSpeechSegmenter* (Doukhan et al. 2018a) to obtain gender correlations. Rezapour et al. (2020) looked at using the iTunes topics and named entities to generate extractive summaries.

Closest to this work, Reddy et al. (2021) analyzed the relationships between linguistic features and engagement (measured via podcast popularity) over the Spotify dataset. In contrast, our work focuses on measuring feature correlations related to *gender* and *discourse*. Hence, we introduce two new modules, the GENDER SEGMENTER module and the CONVERSATIONAL PARSER module, to assist us in our aim of focusing on *gender* and *discourse*, rather than popularity.

Large Language Models (LLM) and Discourse Words. Large language models are trained on gender imbalanced patterns of discourse usage – be it on podcasts, YouTube videos, and/or other social media formats. It is well-recognized that LLMs can inherit and propagate gender stereotypes (Bolukbasi et al. 2016). Thus, with respect to gender, important discourse words should be represented equally in the embedding space, just as stereotype words are via gender debiasing methods (Caliskan, Bryson, and Narayanan 2017; May et al. 2019). Current gender debiasing methods in natural language processing (NLP) typically ignore this prevalent discourse signal, instead focusing on stereotypes, such as occupational stereotypes like *doctor/nurse* (Bolukbasi et al. 2016), and other stereotype categories like *science/arts* and *career/family* in the Word-Embedding Association Test (WEAT) (Caliskan, Bryson, and Narayanan 2017) and the WEAT extension, the Sentence Encoder Association Test (SEAT) (May et al. 2019). Both WEAT and SEAT tests are based on the Implicit Association Test (IAT) from the field of psychology (Greenwald, McGhee, and Schwartz 1998).

While not all discourse words are gender-stereotyped, some are. Consider some of the words from our study: *going*, *know*, and *things* are not stereotyped. However, consider the word *like*, which women in popular media tend to use often – such as in the iconic line from the 2001 hit movie, *Legally Blonde*: *"What, like, it's hard?"* Hence, we differ from WEAT in that we extend WEAT beyond stereotyping, to include discourse words which are correlated with men or women, and hence carry implicit (Seaborn, Chandra, and Fabre 2023; Caliskan, Bryson, and Narayanan 2017; Greenwald, McGhee, and Schwartz 1998; Greenwald and Banaji 1995) gender information and reinforce masculine defaults.

GDCF: Gendered Discourse Correlation Framework (RQ0, RQ1)

In this section, we introduce a framework for discovering gendered discourse words as shown in Figure 1, with features centered around spoken content. We then (RQ0) analyze correlations between between *gender* and *discourse words* to automatically form gendered discourse word lists. Next, (RQ1) we analyze correlations between these gendered discourse words and domains represented by our framework features, and find that, indeed, certain domains (technology/politics, business, and video games) do have discourse-based masculine defaults.

Spotify Podcast Dataset

The Spotify Podcast Dataset is "[t]he largest corpus of transcribed speech data, from a new and understudied domain" (Clifton et al. 2020), which consists of 105,360 podcast episodes in total. The podcasts are randomly sampled from January 1, 2019 to March 1, 2020, and they are "all Spotify owned-and-operated," and contain personally identifiable information and offensive content (Clifton et al. 2020). Each podcast *show* has many *episodes* – analogous to TV shows, where each show has potentially multiple episodes. Clifton et al. (2020) sampled a large collection of podcasts and then filtered out (i) non-English podcasts based on the metadata tags and *langid.py*, a pre-trained multinomial naive Bayes learner (Lui and Baldwin 2011, 2012), (ii) non-professional episodes longer than 90 minutes (hence, this dataset may be biased against beginner/non-professional podcast creators); and (iii) episodes comprised of less than 50% speech.

In our work, we transcribe the podcasts using WhisperX (Bain et al. 2023), a state-of-the-art method based on OpenAI's transformer-based Whisper ASR model, which was trained on 680,000 hours of labeled audio data (Radford et al. 2023) and performs well on discourse-style audio data, as shown in the Appendix in Table 4. (See implementation details in the Appendix). Following Reddy et al. (2021), we apply the following filters to the dataset: (i) truncate episodes to 10 minutes to control for duration; (ii) fil-

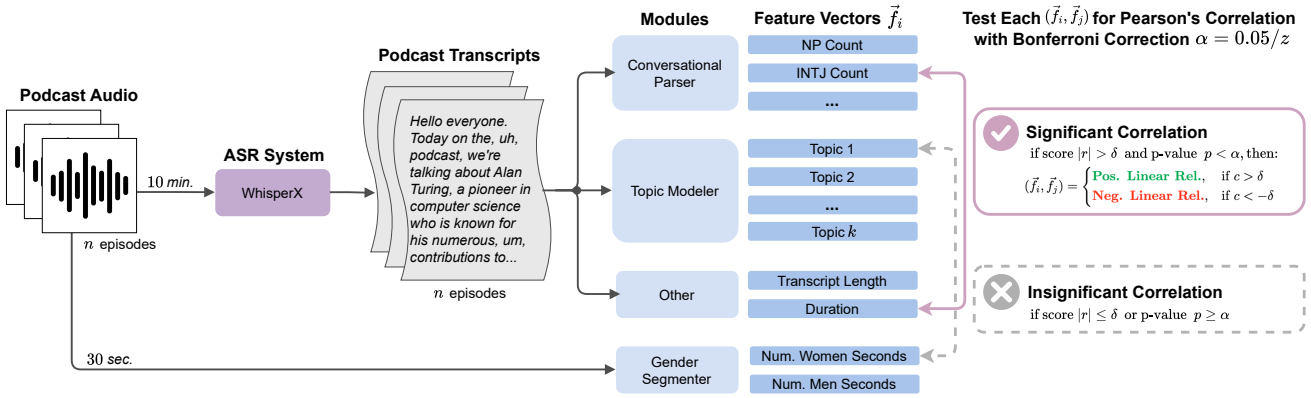

Figure 1: GDCF (Gendered Discourse Correlation Framework) Diagram: Testing for correlations with an example of a significant correlation and an insignificant correlation – all $(\vec{f}_i, \vec{f}_j)$ pairs are labeled *significant* or *insignificant*. $|\vec{f}_i| = 15,117$ podcast episodes. $z = \binom{124}{2} = 7,626$ correlation tests for the 124 total feature vectors.

ter out episodes <10 minutes in duration; (iii) remove the over 3,500 non-English podcasts identified by WhisperX (see Figure 7 in the Appendix) that had bypassed the original Spotify language filter; and (iv) filter out podcasts with less than 10 words.[2] Finally, (v) to control for the impact of a single podcast show having potentially many episodes, we follow Reddy et al. (2021) and only keep one episode per podcast show in our dataset for normalization of the impact of different podcast shows having different numbers of episodes, resulting in 15,117 episodes. We release the correlation results both with and without filter v at: https://www.gendered-discourse.net/extended-results.

Features Definition and Extraction

We define and extract features from the Spotify Podcasts using modules – the CONVERSATIONAL PARSER, the TOPIC MODELER, the GENDER SEGMENTER, and an OTHER module (for the transcript length and duration features). Our features are selected to help us in our aim of characterizing the differences between discourse patterns and topics for men and women. We focus in RQ0 and RQ1 on our most novel result, which is the discovery and significance of *discourse-related masculine defaults*. Results related to the CONVERSATIONAL PARSER and OTHER modules result in the large-scale confirmation of small-scale previous studies, and we provide these results in the Appendix.

TOPIC MODELER MODULE. We perform topic modeling on the podcast transcripts to obtain high-quality *genre* features, as the provided features in the dataset are noisy. For example, the creator-provided descriptions and iTunes categories may be search engine optimized (SEO) to gain more podcast streams – i.e., they may include keywords or other content in an effort to get the podcast ranked higher when users search, rather than to most accurately reflect the content of the podcast. We study two topic models – LDA with non-contextual embeddings, and BERTopic with contextual embeddings:

- **LDA with Non-Contextual Embeddings (Bag-Of-Words).** LDA (Blei, Ng, and Jordan 2003) allows us to represent podcasts as a weighted mixture of multiple topics. The topic representation also models that the same word can be used in different contexts, allowing words to appear under multiple topics. We use NLTK to tokenize the podcast transcripts and CountVectorizer[3] to create the embeddings. We set the seed and run for a maximum of 5 iterations with a batch size of 128. We present the top 10 words for the topics, shown in Table 1 – in the columns *Topic N* and *Topic N Word List*. LDA topics are interpreted by their word list, hence, we manually assign labels to the topics in the column *Topic N Categories*. **Impact of Lemmatization.** We find that lemmatization does not have a significant impact on the output (see the Appendix). Hence, for consistency (Reddy et al. 2021; Clifton et al. 2020; Yang et al. 2019), we do not perform lemmatization. **Impact of Topic Coherence.** We ablate the number of topics in the range $\{40, 60, 80, 100, 120, 140, 160\}$, and find that there is very little variation in topic coherence (in the range of $0.394 - 0.428$) (Röder, Both, and Hinneburg 2015). Following previous work, we use $k = 100$ for our experiments (Reddy et al. 2021; Clifton et al. 2020; Yang et al. 2019).
- **BERTopic with Contextual Embeddings (BERT, ChatGPT, Llama).** For BERTopic (Grootendorst 2022), We conduct these experiments over a subset of $10,000$ randomly-sampled podcasts. BERTopic does not require a set number of topics ahead of time – we find that it pro-

[2] 25 podcasts have no transcribed words at all, despite using a speech filter Clifton et al. (2020). Upon inspection, almost all of these podcasts are ASMR podcast episodes. Similarly, many episodes which have less than 10 words are also ASMR episodes.

[3] We do not use TF-IDF embeddings because TF-IDF scales down the most frequently used terms – and *discourse terms* are high-frequency terms.

duces approx. 50 topics for our set of 10,000 podcasts. We present the results of BERTopic on BERT embeddings, ChatGPT embeddings, and Llama embeddings in Table 2. **Embeddings.** For the BERT embeddings, we use the default SBERT from BERTopic. For the ChatGPT embeddings, we use the `text-embedding-3-large` model from OpenAI. For the Llama embeddings, we obtain the embeddings from `Llama-3.1-8B-Instruct` using the PromptEOL method (Jiang et al. 2023). **Impact of UMAP.** UMAP is a dimensionality-reduction algorithm (McInnes, Healy, and Melville 2018) used in BERTopic to enable the dense, high-dimensional contextual embeddings to be input to HDBSCAN (Campello, Moulavi, and Sander 2013). We experiment with values of $d = \{5, 15, 50, 100\}$ for the reduced dimension, and find that the reduced embeddings are equivalent to each other under these different values of d. Hence, we conduct our experiments with the BERTopic UMAP default value of $d = 5$. We hypotheize that the fine-grained information from the contextual embeddings is lost in the nearest-neighbors approximation via manifold learning from UMAP (McInnes, Healy, and Melville 2018), hence, the output embeddings from UMAP are the same for the 3 different input embeddings.

We examine categories of topics which occur with both the LDA and BERTopic models:[4]

- **Content topics**: The content topics contain words related to content – i.e., content related words for the topic of *yoga* include *energy*, *body*, and *meditation*.
- **Discourse topics**: The discourse words contain words which *are not related to content* – including fillers (e.g., *uh*, *um*), discourse markers (e.g., *well*, *you know*, *I mean*), false starts (e.g., *It was, anyways, I went to Target yesterday*) and more (Merriam-Webster 2024; Shriberg 1994). **These words can indicate differences in the *style of speech*.** Previous works also identify discourse topics (Clifton et al. 2020; Reddy et al. 2021; Yang et al. 2019).

GENDER SEGMENTER MODULE. We use a CNN-based model, *inaSpeechSegmenter*, for state-of-the-art gender detection and segmentation (Doukhan et al. 2018a). This model allows us to analyze and approximate *who*, in terms of gender, is speaking about different content topics, and in what *style* they speak (discourse topics and parts-of-speech). We note that there are exceptions to this method of gender approximation, and discuss this in the Discussion-Limitations section. The model breaks the audio into time segments with five possible values for each segment: *women*, *men*, *music*, *noEnergy*, *noise*. To approximate the gender makeup of each podcast, we run this model on the first 30 seconds, as podcasts often play a short snippet previewing the episode content at the beginning of the podcast and have the hosts introduce the podcast and guests, e.g., $podcast_i$ may have a value of *men seconds* $= 16$, and *women seconds* $= 6$, and *music seconds* $= 8$. **30 Second Approximation.** We test our assumption that the first 30 seconds can approximate the gender makeup of the 10 minute versions of the podcasts. We randomly sample 100 podcasts from the 82k podcasts and run inaSpeechSegmenter (Doukhan et al. 2018a) on the 30 second version and 10 minute version of the podcasts; we then test for significant correlations between these versions: $r(\text{Men}_{\text{30 sec.}}, \text{Men}_{\text{10 min.}}) = 0.79$ and $r(\text{Women}_{\text{30 sec.}}, \text{Women}_{\text{10 min.}}) = 0.82$. Hence, the 30 second labeling is a good approximation. **French-English Language Alignment.** We test the utility of inaSpeechSegmenter for English speech gender identification. We randomly sample 10 podcasts and manually annotate the audio at the seconds-level, and find $r(\text{Men}_{\text{inaSpeechSegmenter}}, \text{Men}_{\text{manual}}) = 0.995$ and $r(\text{Women}_{\text{inaSpeechSegmenter}}, \text{Women}_{\text{manual}}) = 0.981$.

[4]We also note a third category, *language*. For details, see the Appendix.

Feature Correlation Measures

We detail how we test for a significant Pearson's correlation coefficient amongst our feature vectors, $\vec{f_i}$, which were created by our modules – the CONVERSATIONAL PARSER, the TOPIC MODELER, the GENDER SEGMENTER, and an OTHER module (transcript length and duration). We test for a linear relationship between each pair of variables: $H_O : r = 0$, $H_A : r \neq 0$, where H_O is the original hypothesis, H_A is the alternate hypothesis, and r is the Pearson's correlation coefficient. We follow Reddy et al. (2021) and Yang et al. (2019) and apply a Bonferroni correction to our α value of 0.05, setting $\alpha = 0.05/z$, where $z = \binom{124}{2} = 7,626$ for LDA, representing the number of feature relationships we consider. Hence, we reject H_O in favor of H_A if $p \leq \alpha$. Given the largeness of z, our α value becomes small, making our criteria for significance strict and thus suitable for investigating our research questions. Furthermore, we filter our correlations r, such that $\|r\| > 0.1$ for our LDA experiments, and $\|r\| > 0.05$ for our BERTopic experiments (due to the smaller sample size of 10,000 podcasts, and fewer samples may have higher variance). Our results focus on a selection of these significant correlations; the full results are available on the project website: https://www.gendered-discourse.net/extended-results.

RQ0: How are women and men's discourse different?

Using GDCF, our Gendered Discourse Correlation Framework shown in Figure 1, we then analyze significant correlations between between the gender features from the GENDER SEGMENTER module (Doukhan et al. 2018a), and the topic features from the TOPIC MODELER module (Blei, Ng, and Jordan 2003). We use the *discourse topics* to automatically form *gendered discourse word lists* via their significant correlations.

Starting with the first row of Table 1, we see that Topic 3's word list returned by LDA with Non-Contextual Embeddings (Bag-Of-Words) (via the TOPIC MODELER module) contains the words *women, woman, men, baby, pregnant, girls, men, doctor, health, birth* (in descending weighted order). Based on this word list, we manually interpret this topic as being a content topic, specifically about pregnancy, as noted in the column "Topic N Categories." Then, we

Table 1: **LDA with Non-Contextual Embeddings (Bag-Of-Words):** The complete set of significant correlations between gender features and topic features – *both content topics and discourse topics*. Based on r, the Topic N Gender forms the **gendered (discourse) word lists** via Topics 54 and 60 (the masculine word lists) and Topic 62 (the feminine word list).

Topic N	Gender	r	Topic N Word List	Topic N Categories	Topic N Gender
Topic 3	Women	0.15	women, woman, men, baby, pregnant, girls, men, doctor, health, birth	Content - Pregnancy	Women
	Men	-0.14			
Topic 10	Women	0.10	energy, body, feel, mind, space, yoga, love, beautiful, feeling, meditation	Content - Yoga	Women
	Men	-0.12			
Topic 49	Women	-0.21	game, know, think, team, going, mean, play, year, one, good	Content - Sports	Men
	Men	0.17			
Topic 71	Women	0.14	christmas, sex, girl, hair, love, get, date, girls, let, wear	Content - Dating	Women
	Men	-0.14			
Topic 54	Women	–	get, like, know, right, people, going, podcast, make, want, one	Discourse	Men
	Men	0.12			
Topic 60	Women	-0.27	going, know, think, get, got, one, really, good, well, yeah	Discourse	Men
	Men	0.20			
Topic 62	Women	0.33	like, know, really, going, people, want, think, get, things, life	Discourse	Women
	Men	-0.28			

Table 2: **BERTopic with Contextual Embeddings (BERT, ChatGPT, Llama):** The complete set of significant correlations between gender features and topic features for *discourse topics only* (content topics are omitted).

Topic N	Gender	r	Topic N Word List	Topic N Categories	Topic N Gender
Topic 0	Women	-0.08	like, yeah, know, oh, right, podcast, got, going, think, really	Discourse	Men
	Men	0.10			
Topic 2	Women	0.08	life, know, things, really, people, feel, like, want, love, going	Discourse	Women
	Men	-0.08			
Topic 5	Women	0.08	like, know, think, yeah, episode, really, going, anchor, kind, right	Discourse	Women
	Men	–			

look to the gender correlations in the columns "Gender" and "r," and see that $r(\textit{Topic 3, Women}) = +0.15$ and $r(\textit{Topic 3, Men}) = -0.14$. This indicates that the topic of pregnancy positively correlates with women (identified via the GENDER SEGMENTER module), and negatively correlates with men. Therefore, we associate Topic 3 (Content - Pregnancy) with Women, as noted in the "Topic N Gender" column. Similarly, we make these associations in the "Topic N Gender" column for Topics 10, 49, and 71.

Next, we focus on the Topic 54 row. This topic is interpreted using the word list *get, like, know, right, people, going, podcast, make, want, one*. This word list does not refer to any content, hence, we manually interpret this topic as being a discourse topic. Moving to the gender correlations, we see that $r(\textit{Topic 54, Women}) = \emptyset$ and $r(\textit{Topic 3, Men}) = +0.12$. The reason for $r(\textit{Topic 54, Women}) = \emptyset$ is because the correlation between the features *Topic 54* and *Women* did **not** come back as significant. However, due to the positive correlation of 0.12 for *Topic 3* and *Men*, we manually associate *Topic 3* with *Men* in the "Topic N Gender" column. Similarly, we make these associations in the "Topic N Gender" column for Topics 60 and 62. These discourse topics, Topics 54, 60, and 62, and their top-10 word lists then, become our *gendered discourse word lists*. Topics 54 and 60 are associated with men, and hence represent masculine discourse, and Topic 62 is associated with women, and hence represents feminine discourse. We use these LDA word lists for continuity with previous work (Reddy et al. 2021; Clifton et al. 2020; Yang et al. 2019).

Next, looking to Table 2. We perform a smaller-scale (N=10,000 uniformly randomly-selected podcasts) analysis of discourse topics via BERTopic with Contextual Embeddings (BERT, ChatGPT, Llama). As stated previously, we see that the three embeddings all result in the same output topics due to the UMAP dimensionality reduction step in BERTopic, and hypothesize that this is due to the loss of fine-grained information via the nearest-neighbors approximation in UMAP (McInnes, Healy, and Melville 2018). We see that for all three topics – Topic 0, 2, 5 – discourse word lists are formed and have correlations to women, men, or both, similarly to LDA with non-contextual embeddings. Hence, either method can be used depending on the individual application.

RQ1: Are discourse-based masculine defaults present in domain-specific contexts?

Using GDCF, our Gendered Discourse Correlation Framework shown in Figure 1, we analyze correlations between the gendered discourse words discovered in RQ0, and domains represented by topic features. We find that while there may not be a correlation between the gender of the speaker and the domain, there may exist discourse which is more broadly used by speakers of one gender in aggregate.

Starting with the first row of Table 3, we see the masculine discourse topic (Topics 54) and the feminine discourse topic (Topic 62) from RQ0 in the "Topic M" column. Their top-10 word lists are listed in the "Topic M Word List" column. Next, we see a content topic, Topic 11 in the "Topic N" column, and its top-10 word list in the

Table 3: LDA with Non-Contextual Embeddings (Bag-Of-Words): Significant correlations between content topic features and **gendered discourse word lists** (discourse topic features 54, 60, 62, see Table 1) for content topic features which *do not* have direct, significant correlations with gender features, but may broadly be more used by one gender.

Topic N	Topic M	r	Topic N Word List	Topic N Categories	Topic M Word List	Topic M Categories
Topic 11	Topic 54	0.11	data, new, technology, public, bill, theory, science, system, security, article	Content - Technology/ Political	get, like, know, right, people, going, podcast, make, want, one	Discourse (Men)
	Topic 62	-0.20			like, know, really, going, people, want, think, get, things, life	Discourse (Women)
Topic 12	Topic 54	0.24	business, money, company, market, buy, right, million, companies, pay, sell	Content - Business	get, like, know, right, people, going, podcast, make, want, one	Discourse (Men)
Topic 79	Topic 60	0.18	game, games, play, playing, like, played, nintendo, video, fun, switch	Content - Video Games	going, know, think, get, got, one, really, good, well, yeah	Discourse (Men)
	Topic 62	-0.13			like, know, really, going, people, want, think, get, things, life	Discourse (Women)

"Topic N Word List" column: *data, new, technology, public, bill, theory, science, system, security, article*. We manually interpret this topic, then, as being a content topic, specifically about technology/politics, and we note this in the "Topic N Categories" column. Then, looking to the "r" column, we see that r(*Topic 11, Topic 54*) $= +0.11$, and r(*Topic 11, Topic 62*) $= -0.20$. As Topic 54 is a masculine discourse topic, and Topic 62 is a feminine discourse topic, we conclude that the technology/political domain is somewhat dominated by masculine discourse patterns. The use of masculine discourse words in the technology/political domain constitutes a *masculine default* because there is a reward associated with the masculine behavior of using certain discourse words: statistically higher salaries (U.S. Bureau of Labor Statistics 2023). Irrespective of an individual speaker's gender, it is the *use* of these masculine discourse words when discussing technology/politics which constitutes a *masculine default*. These discourse words are, then, also part of the current *technomasculinity* (Bulut 2020). Next, we look to the Topic 12 row. Topic 12 is a content topic which is specifically about business, and r(*Topic 12, Topic 54*) $= +0.24$. Since Topic 54 is a masculine discourse topic, we consider the *reward* associated with the use of masculine discourse words in the business domain, and again, as business results in economic rewards, this is also a *masculine default*. Finally, we look to the Topic 79 row. Topic 79 is a content topic which is specifically about video games, and r(*Topic 79, Topic 60*) $= +0.18$, r(*Topic 79, Topic 62*) $= -0.13$. Since Topic 60 is a masculine discourse topic, we consider the *reward* associated with the use of masculine discourse words in the video game domain, and again, as video games are coupled with computer science (Cheryan and Markus 2020; Cheryan et al. 2013), this association results in economic rewards, making this is also a *masculine default*.

RQ2: Are discourse-based masculine defaults present in LLM embeddings?

Gender differences in LLMs are well-studied (e.g., Dong et al. (2023); Caliskan, Bryson, and Narayanan (2017); May et al. (2019); Bolukbasi et al. (2016)). However, masculine defaults via *gender differences in discourse* in LLMs are not. As LLMs are trained in part over social media, we expect these defaults to be present in the embedding representations of *gendered discourse words*.

Using D-WEAT, our Discourse Word-Embedding Association Test shown in Figure 2, we study the representation of masculine and feminine discourse words in in a state-of-the-art LLM embeddings model from OpenAI, `text-embedding-3-large`. Through hyperparameter studies, we find that the masculine discourse words have a more stable and robust representation, constituting a representational harm (Blodgett et al. 2020) and a masculine default, as this may result in better system outcomes on downstream tasks (Kaneko and Bollegala 2021; Cao et al. 2022).

D-WEAT: Discourse Word-Embedding Association Test

We define a new intrinsic metric, D-WEAT, as an extension of WEAT (Caliskan, Bryson, and Narayanan 2017) which focuses on *gendered discourse words* discovered by our GDCF to "[estimate] fairness in upstream contextualized language representation models" (Cao et al. 2022; Kaneko and Bollegala 2021; Bolukbasi et al. 2016). Similarly to WEAT, we use two sets of target words and two sets of attribute words and measure the association between them.

- **Attribute Words [A_w, A_m]:** We use two parallel word lists from the Word-Embedding Association Test (WEAT) 6B test (Caliskan, Bryson, and Narayanan 2017) to represent the concepts of "men" and "women" in the embedding space – see Figure 2 for an illustration:
  - $A_w = \{$*women, woman, girl, she, her, sister, hers, daughter*$\}$
  - $A_m = \{$*men, man, boy, he, his, brother, him, son*$\}$
- **Target Words [T_w, T_m]:** Target words are words which we "expect to be gender neutral" in the embedding space (Kaneko and Bollegala 2021). WEAT defines target words as gendered category words (e.g., *math* and *poetry*) to study the embedding representation of historically stereotyped subject categories (e.g., *men/math* and *women/poetry*) (Caliskan, Bryson, and Narayanan 2017). We define these words as gendered discourse words (e.g., *going* and *like* from Table 1). This definition allows us to study the embedding representation of gendered discourse words (e.g., *men/going* and *women/like*). **Over-**

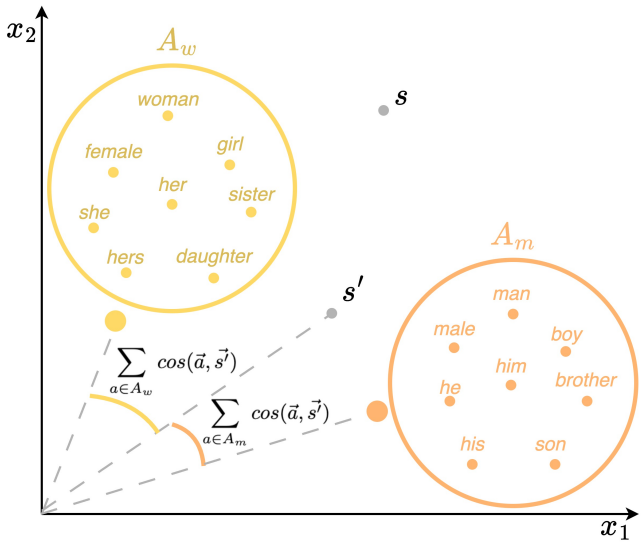

Figure 2: D-WEAT: Plot of the segment vectors $\vec{s}$ and $\vec{s'}$, and the word vectors, $\vec{w} \in A_w$, and $\vec{w} \in A_m$, projected into a two-dimensional space for illustrative purposes. The cosine similarity for s' and A_w, and s' and A_m is depicted; the cosine similarity for s and A_w, and s and A_m is calculated in the same way.

lapping Words. Our topics which represent gendered discourse, Topic 60 (masculine discourse, forms T_m) and Topic 62 (feminine discourse, forms T_w), contain some overlapping words, hence, we apply the following rules to arrive at our two lists: (i) if there is a word in the same position in both lists, it is removed, (ii) if there is a word in different positions in both lists, it remains in the list where it occurs first. Thus, we form T_w and T_m using our discourse topics from LDA for consistency with previous work:

- $T_w = \{$*like, really, people, want, things, life, feel, time, something, right*$\}$ These are the top weighted words *post-filtering* from Topic 62 (Table 1), which is significantly positively correlated with *women* and negatively correlated with *men*, representing the feminine discourse style.
- $T_m = \{$*going, think, get, got, one, good, well, yeah, bit, week*$\}$ These are the top 10 weighted words *post-filtering* from Topic 60 (Table 1),[5] which is significantly positively correlated with *men* and negatively correlated with *women*, representing the masculine discourse style.

We also experiment with the gendered discourse topics discovered via BERTopic, and see a similar result. See the Appendix for details.

[5]Topic 54 is only positively correlated with *men*, and has no significant correlation with *women*; thus, we use Topic 60, as it has significant correlations for both *men* (+) and *women* (−).

Dataset Formation [$(s, s') \in S_w$, $(s, s') \in S_m$]. Then, we form two sets of samples: S_w is the *women* set of samples and S_m is the *men* set of samples from the Spotify Podcast Dataset, as determined by t_{women} and t_{men}, features for the total number of men or women seconds as determined by the GENDER SEGMENTER module. For $(s, s') \in S_w$, $t_{women} \geq \tau$ and for $(s, s') \in S_m$, $t_{men} \geq \tau$, where τ is a parameter we vary to control the minimum number of men or women seconds. We search for $\tau \in \{20, 25, 30\}$, and study the impact of this parameter in Figures 3 and 4. We form S_w and S_m by (i) sampling 100 podcasts which meet the τ threshold, (ii) taking 3 segments from each of those podcasts – each of these segments is 3 sentences long,[6] (iii) keeping segments which have at least γ number of words from T_w for S_w and T_m for S_m. Consequently, fewer segments are kept as γ increases. For each segment $s \in S_m$, we form s' by replacing each of the words in T_m with a randomly-selected word from T_w. For example, if:

$s =$ *And I was* ***going****, hey, it's cold outside...*

Then the word *going*$\in T_m$ is replaced with a randomly-selected word from T_w to form s':

$s' =$ *And I was* ***like****, hey, it's cold outside...*

Hence, $\gamma = 1$ for s and s', because there is only one word, *going*, which is in T_m to replace with a randomly-selected word from T_w. The process is similar for each sample $s \in S_w$: we form s' by replacing each of the words in T_w with a randomly-selected word from T_m. This process simulates keeping the sentence the same, except for the gendered discourse words. In future work, a more nuanced approach would involve context-aware discourse word replacement, rather than random replacement.

LLM Representation. We obtain the embedding representation of s and s' from using contextual embeddings. Specifically, the Open AI embedding model `text-embedding-3-large`[7] due to the popularity of the OpenAI models. We have a similar finding with Llama embeddings, and report the results in the Appendix. The cosine similarity between these two embedding vectors is $cos(\vec{s}, \vec{s'}) = \vec{s}^T \cdot \vec{s'} / \|\vec{s}\| \|\vec{s'}\|$. We conventionally use the terms *similarity* and *distance* interchangeably – the more similar two vectors are, the closer they are. We expect cosine similarity to be equivalent if we flip the discourse words, assuming these words are ungendered. If they move in such a way that they are more similar to either the *men* or *women* concepts in the embedding space, then that means that these words carry gender information.

While non-contextual embeddings were useful for topic modeling via LDA – because LDA is designed to work on such embeddings – they are not a good choice for this experiment. First, count embeddings are not a suitable choice because the discourse words are high-frequency, and therefore they dominate the cosine calculation and wash away the lower-frequency more informative words with this representation. Second, TF-IDF embeddings are again not a

[6]https://www.nltk.org/api/nltk.tokenize.html.

[7]https://platform.openai.com/docs/guides/embeddings

good fit for modeling discourse words because discourse words are high-frequency (similar to stopwords), and TF-IDF is specifically designed to focus on low-frequency terms. Hence, we use dense contextual embeddings for our experiments to retain meaning from low-frequency words and examine the context and nuance of the high-frequency discourse words.

Measuring Movement via Women % and Men % . To measure the movement in the embedding space, we calculate Δ_w and Δ_m for each (s, s') pair in S_w and S_m. The Δ_w and Δ_m indicate how s' moves in relation to the *women* and *men* concepts, A_w and A_m, in the embedding space, when the discourse words are replaced. Specifically, as illustrated in Figure 2, we sum the total cosine similarity between s', and each of the words $w \in A_w$, and $w \in A_m$. We do the same for s.[8] Then we calculate the movement by taking the difference of summed cosine similarity values, as shown in Equations 1 and 2:

$$\Delta_w = \sum_{a \in A_w} cos(\vec{a}, \vec{s'}) - \sum_{a \in A_w} cos(\vec{a}, \vec{s}) \quad (1)$$

$$\Delta_m = \sum_{a \in A_m} cos(\vec{a}, \vec{s'}) - \sum_{a \in A_m} cos(\vec{a}, \vec{s}) \quad (2)$$

We then use two counter variables, C_m and C_w, to indicate: How does s' move in relation to s, and the *women* and *men* concepts – A_w and A_m? **Which concept – women (A_w) or men (A_m) – did s move closer to, when the discourse words were replaced to form s'?**

$$C_w = C_w + 1 \begin{cases} \text{if } \Delta_m, \Delta_w > 0 \text{ and } \Delta_w > \Delta_m \\ \text{if } \Delta_m, \Delta_w < 0 \text{ and } \Delta_w < \Delta_m \\ \text{else if } \Delta_w > \Delta_m \end{cases} \quad (3)$$

$$C_m = C_m + 1 \begin{cases} \text{if } \Delta_m, \Delta_w > 0 \text{ and } \Delta_m > \Delta_w \\ \text{if } \Delta_m, \Delta_w < 0 \text{ and } \Delta_m < \Delta_w \\ \text{else if } \Delta_m > \Delta_w \end{cases} \quad (4)$$

As shown in Equations 3 and 4, there are three possible situations in which the movement from $\vec{s}$ to $\vec{s'}$ can occur – $\vec{s'}$ moves closer to both the *women* and *men* concepts, $\vec{s'}$ moves farther from both the *women* and *men* concepts, and $\vec{s'}$ moves closer to one concept and farther from the other:

1. For $\Delta_m, \Delta_w > 0$, s' **moves closer** to both A_w and A_m – the *women* and *men* concepts. In this case, whichever concept s' moves closer to gets its corresponding counter, C_w and C_m, incremented.
2. For $\Delta_m, \Delta_w < 0$, s' **moves farther** from both A_w and A_m – the *women* and *men* concepts. In this case, whichever concept s' moves less far from gets its corresponding counter, C_w and C_m, incremented.

[8] We take the average over 3 calculations of the sum of the cosine similarity, to account for the small variation in the embeddings returned by the Open AI Embeddings API.

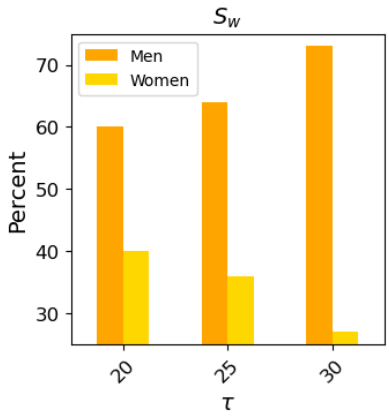

Figure 3: Impact of τ on the average percentage of S_w segments which move closer to the *women* concept (A_w) versus the *men* (A_m) concept.

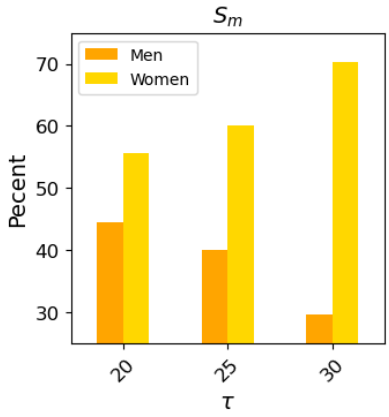

Figure 4: Impact of τ on the average percentage of S_m segments which move closer to the *women* concept (A_w) versus the *men* (A_m) concept.

3. For the final case, either $\Delta_m > 0$ or $\Delta_w > 0$, while the other is < 0, so s' **moves closer** to either A_w and A_m **and moves farther from the other**. In this case, whichever concept s' moves closer to gets its corresponding counter, C_w and C_m, incremented.

We obtain C_w and C_m counts and take the average. We then report these counts, C_w and C_m, in Figures 3, 4, 5, and 6 in terms of percentages, to normalize for the impact of γ, as there are fewer total samples which have $\geq \gamma$ discourse words to swap as γ increases.

Impact of τ. We study the impact of varying the minimum number of women or men seconds τ in $\{20, 25, 30\}$. For this analysis, we set $\gamma = 6$, a middle value in our γ ablation – neither having the most number of samples, nor the least.

Starting with Figure 3, for S_w, we see that at $\tau = 20$, the men percent is 60% and the women percent is 40%. Moving along the x-axis, we see that the men percentage continues to increase, reaching a maximum of $\approx 73\%$ at $\tau = 30$,

while the women percentage continues to decrease, reaching a minimum of $\approx 27\%$ also at $\tau = 30$. In Figure 4, for S_m, we see that at $\tau = 20$, the man percent is $\approx 55\%$ and the woman percent is $\approx 45\%$. Then, at $\tau = 25$, the gap between the man percent and woman percent widens. This gap widens again at $\tau = 30$, reaching an extreme with a man percent of $\approx 70\%$, and a woman percent of $\approx 30\%$. Comparing Figures 3 and 4, we see that the initial gap at $\tau = 20$ is larger on S_w (60%/40%) versus S_m (55%/45%). This trend continues through $\tau = 30$, where the gap is larger on S_w (73%/27%) versus S_m (70%/30%). Therefore, because the masculine discourse words have a more stable representation (this is a representational harm (Blodgett et al. 2020)) as compared to feminine discourse words, the LLM can obtain better performance on downstream tasks (Cao et al. 2022; Kaneko and Bollegala 2021) (this is the *reward)* in the presence of men discourse words, and this constitutes a masculine default.

Impact of γ. We study the impact of varying the minimum number of swaps, γ, in the range $\{1, 2, ..., 10\}$ for each segment for the men segments and the women segments. For this analysis, we set $\tau = 30$, as this is the value for which the gap is the greatest for the men percentage and the women percentage for S_w and S_m.

Starting with Figure 5, at $\gamma = 1$ along the x-axis, we see that the men percent is approx. 60%, while the women percent is approx. 40%. This means that on average, when the discourse words from T_w in each sample $s \in S_w$ were replaced with the randomly-selected discourse words from T_m, $\vec{s'}$ moved closer to the man concept (A_m) than the women concept (A_m) in the contextual embedding space. Moving along the x-axis, as γ increases, this gap widens, reaching an extreme at $\gamma = 10$ of approx. 90% for the men percentage and 10% for the women percentage. This indicates that the embedding model does indeed learn gendered patterns of discourse. Similarly, as shown in Figure 6, the gap between the men and women percent increases moving along the x-axis as γ increases, with the women percent dropping to 0% at the extreme values of $\gamma = 9, 10$. Comparing Figures 5 and 6, we see that the gap for $\gamma = 1$ in Figure 5 is wider than the gap for $\gamma = 1$ in Figure 6, indicating that masculine discourse has a more robust representation in the embedding space. This is because with the same number of discourse word replacements, γ, more segments, on average, from the S_w segments move closer to the A_m man concept in the embedding space – approx. 60% – while, on average, only approx. 55% of the segments on average from the S_m segments move closer to the A_w women concept in the embedding space. We interpret this imbalance in percent (wider gap for S_w men percent than S_m women percent) as evidence of masculine defaults being learned by, and thus ingrained in, this widely-used embedding model. (Note that we focus on the $\gamma = 1-6$ segments as these values of γ have lots of samples, whereas the extremes of $\gamma = 7-10$ have much fewer samples and therefore higher variance, hence at these values we can interpret with less specificity that the gap between the men percentage and women percentage tends to increase.) This imbalance in percent is an issue as "such biases can easily propagate to the downstream NLP applications that use contextualised text embeddings" (Kaneko and Bollegala 2021; Bolukbasi et al. 2016).

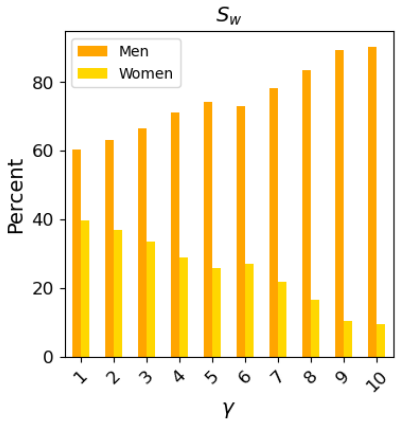

Figure 5: Impact of γ on the average percentage of S_w segments which move closer to the women concept (A_w) versus the men (A_m) concept.

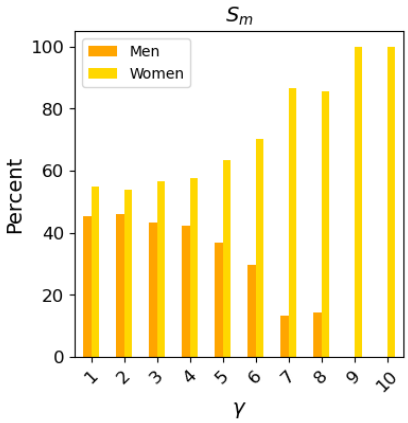

Figure 6: Impact of γ on the average percentage of S_m segments which move closer to the women concept (A_w) versus the men (A_m) concept.

Discussion

In this paper we analyzed the *discourse-based masculine defaults*. We proposed a framework with two parts (i) the *Gendered Discourse Correlation Framework (GDCF)*, a framework for identifying and analyzing gendered discourse; and (ii) *Discourse Word-Embedding Association Test (D-WEAT)*, a measure of the gender bias associated with gendered discourse words in LLMs. We studied 15,117 podcast episodes from the Spotify Podcast Dataset (Clifton et al. 2020), and used GDCF to automatically form *gendered discourse word lists*.

Limitations

Different forms of media – such as short videos that are prevalent on TikTok, Instagram, and YouTube, long videos on YouTube, streamers on Twitch, and even text-based media such as posts on Facebook, X (Twitter), and Instagram – may have different language patterns and styles, and further work can explore gendered discourse in these contexts. Additionally, Spotify Podcasts are produced for people who can listen to them. This imposes a socioeconomic constraint on the podcast data, as listeners must likely own an electronic device – usually a mobile device (NPR and Research 2023). In terms of creators, Spotify estimates that approx. 10% of the podcasts in the dataset were made by professional creators, while the remaining 90% were made by amateur creators (Clifton et al. 2020). This long tail of amateur creators likens podcasts to social media, as Spotify for Podcasters (previously Anchor) makes it easy to produce podcasts – i.e., producing a podcast does not necessarily require professional equipment. We lack access to certain demographic metadata in the Spotify Podcasts Dataset, including gender, age, and socioeconomic status. Our findings should not be generalized to content which is in a different format – however, GDCF can be used to analyze this content. The Spotify data we use is limited to English speech, and 67% originates from the US (as indicated by creator tags on 2,223 of the podcasts) (Clifton et al. 2020).

Our study has a key limitation in that we utilize the binary gender definition,[1] rather than treating gender as a spectrum or otherwise modeling nonbinary genders. This lack of representation results in lowered capabilities of GDCF and D-WEAT, and hence, more nuanced approaches should be developed by future work in order to create inclusive representations of gender. Using *inaSpeechSegmenter* (Doukhan et al. 2018a) in the GENDER SEGMENTER MODULE, we approximate gender via sex – however, we note that there are many exceptions to this approximation. For example: persons with high or low voices for their sex, intersex people, transgender people, and more. While this tool, and hence this approximation of gender via sex, has been successfully used previously (i.e., Doukhan et al. (2018b); Martikainen, Karlgren, and Truong (2022)), we do not claim that this is a perfect approximation, simply that it is the best one that we have at this time based on the data and model available to us towards building a discourse-based debiasing method. Future work should explore a more representative gender feature extraction step. We view this research gap as an opportunity for speech and social media researchers to create datasets to enable this technology. Rather than using binary gender, speech audio datasets can include a metadata field for self-identified genders.

Theoretical Implications

First, the use of gendered discourse words can be considered a type of *gender performativity* (Butler 1988, 2009; West and Zimmerman 1987; Unger 1979; Muehlenhard and Peterson 2011), wherein the discourse words are part of a *gender schema* (Bem 1984; West and Zimmerman 1987). Hence, we identify specific words which are part of the current *hegemonic masculine* strategy (Connell 1995, 1987) – and in the domain of technology, discourse words which are part of the *technomasculine* strategy (Cooper 2000; Lockhart 2015; Bulut 2020). There exist rewards for the use of masculine discourse words in the following ways: in the domains of technology/politics and business this language is rewarded with economic rewards, and in LLMs, this language is rewarded with a more stable representation. Hence, these gendered discourse words constitute a *masculine default* (Cheryan and Markus 2020), and we contribute this framework, GDCF, for the discovery and analysis of gendered discourse words.

Second, D-WEAT is an intrinsic metric which can be used to debias LLMs, similarly to WEAT (Caliskan, Bryson, and Narayanan 2017), and the inclusion of discourse words broadens the debiasing task in natural language processing. We focus in this work on measuring *intrinsic bias*. An important future direction includes studying gendered discourse words in the context of *extrinsic bias* (Blodgett et al. 2020), as indicated by these findings from Cao et al. (2022):"[W]e find that correlations between intrinsic and extrinsic metrics are sensitive to alignment in notions of bias, quality of testing data, and protected groups. We also find that extrinsic metrics are sensitive to variations on experiment configurations, such as to classifiers used in computing evaluation metrics. Practitioners thus should ensure that evaluation datasets correctly probe for the notions of bias being measured." Hence, analyzing bias at the intrinsic and extrinsic level are two separate problems which are both important, and future work can consider extrinsic debiasing w.r.t. gendered discourse words.

Policy Implications

Policymakers – in government or platforms such as Spotify – could implement measures by which to mitigate bias in LLMs w.r.t. gender. Specifically, policymakers could regulate the use of D-WEAT to impose an unbiased representation of discourse words with respect to gender. D-WEAT could be run regularly, and a threshold could be set to determine what an "acceptable" level of bias is in a given LLM. Broadly, D-WEAT can join *a set of debiasing methods, tools, and datasets* (Bolukbasi et al. 2016; Caliskan, Bryson, and Narayanan 2017; May et al. 2019; Nangia et al. 2020; Nadeem, Bethke, and Reddy 2020; Guo, Yang, and Abbasi 2022; He et al. 2022; Cheng, Durmus, and Jurafsky 2023; Dong et al. 2023) which can be employed to regulate bias in LLMs.

Ethical Implications

A potential ethical concern is that tools used to remove bias can also be used to exacerbate bias. GDCF and D-WEAT could potentially be used to discover discourse words in audio-text corpora, and then *increase* the gender bias of the LLM embeddings. This abuse of the framework would be a *representational harm* (Blodgett et al. 2020). However, a more important point is that it is hard to undo bias issues without knowing how that bias manifests; here, we provide a framework to identify and quantify this subtle gender bias so that it can be undone in powerful LLMs.

Acknowledgments

We would like to thank Majid Alfifi for the discussion and the help transcribing the podcasts and David Jefts for the discussion.

References

American Society for Engineering Education. 2022. Engineering Engineering Technology By The Numbers.

Bain, M.; Huh, J.; Han, T.; et al. 2023. WhisperX: Time-Accurate Speech Transcription of Long-Form Audio. *INTERSPEECH 2023*.

Beattie, G. W.; and Butterworth, B. L. 1979. Contextual Probability and Word Frequency as Determinants of Pauses and Errors in Spontaneous Speech. *Language and Speech*, 22(3): 201–211.

Beauvoir, S. d. 1949. *The Second Sex*.

Bem, S. L. 1984. Androgyny and gender schema theory: a conceptual and empirical integration. *Nebraska Symposium on Motivation. Nebraska Symposium on Motivation*, 32: 179–226.

Blei, D. M.; Ng, A. Y.; and Jordan, M. I. 2003. Latent Dirichlet Allocation. *Journal of Machine Learning Research*, 3(Jan): 993–1022.

Blodgett, S. L.; Barocas, S.; III, H. D.; and Wallach, H. 2020. Language (Technology) is Power: A Critical Survey of "Bias" in NLP. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 5454–5476.

Bolukbasi, T.; Chang, K.-W.; Zou, J.; Saligrama, V.; and Kalai, A. 2016. Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings. *arXiv*.

Bortfeld, H.; Leon, S. D.; Bloom, J. E.; et al. 2001. Disfluency Rates in Conversation: Effects of Age, Relationship, Topic, Role, and Gender. *Language and Speech*, 44(2): 123–147.

Branigan, H.; Lickley, R.; and McKelvie, D. 1999. Non-Linguistic Influences on Rates of Disfluency in Spontaneous Speech. In *Conference of Phonetic Sciences*, 387–390.

Brennan, S. E.; and Schober, M. F. 2001. How Listeners Compensate for Disfluencies in Spontaneous Speech. *Journal of Memory and Language*, 44(2): 274–296.

Bulut, E. 2020. *A Precarious Game*. Ithaca, NY: Cornell University Press. ISBN 9781501746543.

Butler, J. 1988. Performative Acts and Gender Constitution An Essay in Phenomenology and Feminist Theory. *Theatre Journal*, 40(4): 519.

Butler, J. 2009. Performativity, precarity and sexual politics. *AIBR. Revista de Antropología Iberoamericana*, 4(3).

Caliskan, A.; Bryson, J. J.; and Narayanan, A. 2017. Semantics derived automatically from language corpora contain human-like biases. *Science*, 356(6334): 183–186.

Campello, R. J.; Moulavi, D.; and Sander, J. 2013. Density-based clustering based on hierarchical density estimates. In *Pacific-Asia conference on knowledge discovery and data mining*, 160–172. Springer.

Cao, Y.; Pruksachatkun, Y.; Chang, K.-W.; Gupta, R.; Kumar, V.; Dhamala, J.; and Galstyan, A. 2022. On the Intrinsic and Extrinsic Fairness Evaluation Metrics for Contextualized Language Representations. *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, 561–570.

Charniak, E.; and Johnson, M. 2001. Edit Detection and Parsing for Transcribed Speech. In *NAACL*.

Cheng, M.; Durmus, E.; and Jurafsky, D. 2023. Marked Personas: Using Natural Language Prompts to Measure Stereotypes in Language Models. *arXiv*.

Cheryan, S.; and Markus, H. R. 2020. Masculine defaults: Identifying and mitigating hidden cultural biases. *Psychological Review*, 127(6): 1022–1052.

Cheryan, S.; Plaut, V. C.; Handron, C.; and Hudson, L. 2013. The Stereotypical Computer Scientist: Gendered Media Representations as a Barrier to Inclusion for Women. *Sex Roles*, 69(1–2): 58–71.

Clark, H. H.; and Tree, J. E. F. 2002. Using Uh and Um in Spontaneous Speaking. *Cognition*, 84(1): 73–111.

Clifton, A.; Reddy, S.; Yu, Y.; et al. 2020. 100,000 Podcasts: A Spoken English Document Corpus. In *COLING*, 5903–5917.

Connell, R. 1987. *Gender and power: society, the person, and sexual politics*. Stanford University Press.

Connell, R. 1995. *Masculinities*. Allen Unwin.

Cooper, M. 2000. Being the "Go-To Guy": Fatherhood, Masculinity, and the Organization of Work in Silicon Valley. *Qualitative Sociology*, 23(4): 379–405.

Corley, M.; MacGregor, L. J.; and Donaldson, D. I. 2007. It's the Way That You, Er, Say It: Hesitations in Speech Affect Language Comprehension. *Cognition*, 105(3): 658–668.

Corley, M.; and Stewart, O. W. 2008. Hesitation Disfluencies in Spontaneous Speech: The Meaning of Um. *Language and Linguistics Compass*, 2(4): 589–602.

Diachek, E.; and Brown-Schmidt, S. 2023. The Effect of Disfluency on Memory For What Was Said. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 49(8): 1306.

Dong, X.; Zhu, Z.; Wang, Z.; Teleki, M.; and Caverlee, J. 2023. Co^2PT: Mitigating Bias in Pre-trained Language Models through Counterfactual Contrastive Prompt Tuning. In Bouamor, H.; Pino, J.; and Bali, K., eds., *Findings of the Association for Computational Linguistics: EMNLP 2023*, 5859–5871. Association for Computational Linguistics.

Doukhan, D.; Carrive, J.; Vallet, F.; et al. 2018a. An Open-Source Speaker Gender Detection Framework for Monitoring Gender Equality. In *ICASSP*. IEEE.

Doukhan, D.; Poels, G.; Rezgui, Z.; et al. 2018b. Describing Gender Equality in French Audiovisual Streams with a Deep Learning Approach. *VIEW Journal of European Television History and Culture*, 7(14): 103–122.

Ghai, B.; Hoque, M. N.; and Mueller, K. 2021. Wordbias: An interactive visual tool for discovering intersectional biases encoded in word embeddings. In *Extended Abstracts of*

the 2021 CHI Conference on Human Factors in Computing Systems, 1–7.

Godfrey, J. J.; and Holliman, E. 1997. Switchboard-1 Release 2. *Linguistic Data Consortium*.

Google Cloud. 2023. Speech-To-Text.

Goree, S.; Crandall, D.; and Su, N. M. 2023. "It Was Really All About Books:" Speech-like Techno-Masculinity in the Rhetoric of Dot-Com Era Web Design Books. *ACM Transactions on Computer-Human Interaction*, 30(2): 1–27.

Greenwald, A. G.; and Banaji, M. R. 1995. Implicit Social Cognition: Attitudes, Self-Esteem, and Stereotypes. *Psychological Review*, 102(1): 4–27.

Greenwald, A. G.; McGhee, D. E.; and Schwartz, J. L. K. 1998. Measuring Individual Differences in Implicit Cognition: The Implicit Association Test. *Journal of Personality and Social Psychology*, 74(6): 1464–1480.

Grootendorst, M. 2022. BERTopic: Neural topic modeling with a class-based TF-IDF procedure. *arXiv preprint arXiv:2203.05794*.

Guo, Y.; Yang, Y.; and Abbasi, A. 2022. Auto-Debias: Debiasing Masked Language Models with Automated Biased Prompts. *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 1012–1023.

He, J.; Xia, M.; Fellbaum, C.; and Chen, D. 2022. MABEL: Attenuating Gender Bias using Textual Entailment Data. *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, 9681–9702.

Honnibal, M.; and Johnson, M. 2014. Joint Incremental Disfluency Detection and Dependency Parsing. *TACL*.

James, D.; and Drakich, J. 1993. Understanding Gender Differences in Amount of Talk: A Critical Review of Research.

Jamshid Lou, P.; and Johnson, M. 2020. Improving Disfluency Detection by Self-Training a Self-Attentive Model. In *ACL*.

Jiang, T.; Huang, S.; Luan, Z.; Wang, D.; and Zhuang, F. 2023. Scaling Sentence Embeddings with Large Language Models. *arXiv*.

Johnson, I.; Lemmerich, F.; Sáez-Trumper, D.; West, R.; Strohmaier, M.; and Zia, L. 2021. Global gender differences in Wikipedia readership. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 15, 254–265.

Johnson, M.; and Charniak, E. 2004. A TAG-Based Noisy-Channel Model of Speech Repairs. In *ACL*.

Kalhor, G.; Gardner, H.; Weber, I.; and Kashyap, R. 2023. Gender Gaps in Online Social Connectivity, Promotion and Relocation Reports on LinkedIn. In *Proceedings of the International AAAI Conference on Web and Social Media*.

Kaneko, M.; and Bollegala, D. 2021. Debiasing Pre-trained Contextualised Embeddings. *arXiv*.

Leaper, C.; and Ayres, M. M. 2007. A Meta-Analytic Review of Gender Variations in Adults' Language Use: Talkativeness, Affiliative Speech, and Assertive Speech. *Personality and Social Psychology Review*, 11(4): 328–363.

Linguistic Data Consortium. 2002. 2000 HUB5 English Evaluation Transcripts.

Lockhart, E. A. 2015. *Nerd/Geek masculinity: Technocracy, Rationality, and gender in nerd culture's countermasculine hegemony*. Ph.D. thesis.

Lou, P. J.; Wang, Y.; and Johnson, M. 2019. Neural Constituency Parsing of Speech Transcripts. *arXiv preprint arXiv:1904.08535*.

Lui, M.; and Baldwin, T. 2011. Cross-domain Feature Selection for Language Identification. In *IJCNLP*.

Lui, M.; and Baldwin, T. 2012. langid.py: An Off-the-Shelf Language Identification Tool. In *ACL 2012 System Demonstrations*.

Martikainen, K.; Karlgren, J.; and Truong, K. P. 2022. Exploring Audio-Based Stylistic Variation in Podcasts. In *INTERSPEECH 2022*.

May, C.; Wang, A.; Bordia, S.; Bowman, S. R.; and Rudinger, R. 2019. On Measuring Social Biases in Sentence Encoders. *arXiv*.

McInnes, L.; Healy, J.; and Melville, J. 2018. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*.

Merriam-Webster. 2024. Discourse.

Mitchell, M.; Santorini, B.; Marcinkiewicz, M.; et al. 1999. Treebank-3 LDC99T42 Web Download. *Linguistic Data Consortium*, 3: 2.

Muehlenhard, C. L.; and Peterson, Z. D. 2011. Distinguishing Between Sex and Gender: History, Current Conceptualizations, and Implications. *Sex Roles*, 64(11–12): 791–803.

Nadeem, M.; Bethke, A.; and Reddy, S. 2020. StereoSet: Measuring stereotypical bias in pretrained language models. *arXiv*.

Nangia, N.; Vania, C.; Bhalerao, R.; and Bowman, S. R. 2020. CrowS-Pairs: A Challenge Dataset for Measuring Social Biases in Masked Language Models. *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 1953–1967.

NPR; and Research, E. 2023. The Spoken Word Audio Report.

Ovalle, A.; Goyal, P.; Dhamala, J.; Jaggers, Z.; Chang, K.-W.; Galstyan, A.; Zemel, R.; and Gupta, R. 2023. "I'm fully who I am": Towards Centering Transgender and Non-Binary Voices to Measure Biases in Open Language Generation. *2023 ACM Conference on Fairness, Accountability, and Transparency*, 1246–1266.

Panayotov, V.; Chen, G.; Povey, D.; et al. 2015. Librispeech: An ASR Corpus Based On Public Domain Audio Books. In *ICASSP*.

Radford, A.; Kim, J. W.; Xu, T.; et al. 2023. Robust Speech Recognition via Large-Scale Weak Supervision. In *ICML*.

Reddy, S.; Lazarova, M.; Yu, Y.; et al. 2021. Modeling Language Usage and Listener Engagement in Podcasts. In *ACL*.

Rezapour, R.; Reddy, S.; Clifton, A.; et al. 2020. Spotify at TREC 2020: Genre-Aware Abstractive Podcast Summarization. In *TREC*, NIST Special Publication. National Institute of Standards and Technology (NIST).

Röder, M.; Both, A.; and Hinneburg, A. 2015. Exploring the space of topic coherence measures. In *Proceedings of the eighth ACM international conference on Web search and data mining*, 399–408.

Sczesny, S.; Formanowicz, M.; and Moser, F. 2016. Can Gender-Fair Language Reduce Gender Stereotyping and Discrimination? *Frontiers in Psychology*, 7: 25.

Seaborn, K.; Chandra, S.; and Fabre, T. 2023. Transcending the "Male Code": Implicit Masculine Biases in NLP Contexts. *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, 1–19.

Shriberg, E. 1996. Disfluencies In Switchboard. In *International Conference on Spoken Language Processing*.

Shriberg, E. E. 1994. *Preliminaries to a Theory of Speech Disfluencies*. Ph.D. thesis.

Taylor, A.; Marcus, M.; and Santorini, B. 2003. The Penn Treebank: An Overview. *Treebanks: Building and Using Parsed Corpora*.

The Pew Research Center, C. S. A. 2023. Audio and Podcasting Fact Sheet.

Tran, T.; Toshniwal, S.; Bansal, M.; et al. 2017. Parsing Speech: a Neural Approach to Integrating Lexical and Acoustic-Prosodic Information. *arXiv preprint arXiv:1704.07287*.

Tree, J. E. F. 1995. The Effects of False Starts and Repetitions on the Processing of Subsequent Words in Spontaneous Speech. *Journal of Memory and Language*, 34(6): 709–738.

Unger, R. K. 1979. Toward a redefinition of sex and gender. *American Psychologist*, 34(11): 1085–1094.

U.S. Bureau of Labor Statistics. 2023. Occupational Employment and Wage Statistics.

Valero, F. B.; Baranes, M.; and Epure, E. V. 2022. Topic Modeling on Podcast Short-Text Metadata. In *ECIR*.

Wang, A.; Pappu, A.; and Cramer, H. 2021. Representation of Music Creators on Wikipedia, Differences in Gender and Genre. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 15, 764–775.

Wang, Y.; and Horvát, E.-Á. 2019. Gender differences in the global music industry: Evidence from musicbrainz and the echo nest. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 13, 517–526.

West, C.; and Zimmerman, D. 1987. Doing Gender. *Gender and Society*, 1: 125–151.

Yang, L.; Wang, Y.; Dunne, D.; et al. 2019. More Than Just Words: Modeling Non-Textual Characteristics of Podcasts. In *WSDM*.

Appendix

We release the code at https://github.com/mariateleki/masculine-defaults and the extended results at the project website: https://www.gendered-discourse.net/extended-results.

WhisperX

Transcription Details with WhisperX. We re-transcribe the podcasts using WhisperX (Bain et al. 2023), a state-of-the-art method based on OpenAI's transformer-based Whisper ASR model, which was trained on 680,000 hours of labeled audio data (Radford et al. 2023). WhisperX speeds up Whisper transcription by 12x using Voice Activity Detection to detect active speech regions, and then a cut and merge strategy to allow for parallel batch transcription with Whisper. We transcribe the podcast audio using a batch size of 24. We use *large-v2* for our model. The models were run across 8 machines with 60 NVIDIA RTX A4000 GPUs and 2 NVIDIA RTX A5000 GPUs. It took about a day and a half to transcribe all the podcast audio.

Comparing Google ASR to WhisperX in Terms of Disfluent Token Transcription. As shown in Table 4, we conduct a small-scale experiment on 100 randomly-selected podcasts to evaluate the transcription quality of WhisperX as compared to Google ASR (Google Cloud 2023) (which was originally used to transcribe the dataset by Clifton et al. (2020)) in terms of disfluent token transcription. We fix the seeds for each random sample of size 100, and run the sampling 5 times.

Starting with the first row of Table 4, we see the mean and standard deviation of the number of *uh* tokens present in the transcripts. WhisperX transcribed on average 1.07 *uh* tokens per podcast, while Google ASR transcribed on average 0.18 *uh* tokens per podcast for the same podcasts. Similarly for *um* and *well*, the average number of *um* tokens transcribed is higher for WhisperX than Google ASR. We notice similar standard deviations across WhisperX and Google ASR for the *well* token, indicating consistency in the transcriptions. Thus, we hypothesize that the vast amount of training data used to train Whisper (Radford et al. 2023) contained *um* and *uh* tokens, and therefore WhisperX is able to transcribe these common disfluent tokens, whilst Google ASR is less capable of transcribing these common disfluent tokens.

The large standard deviation values may be due to the heterogeneity of the podcasts, as some are scripted and likely contain less disfluent tokens, while others are unscripted and may contain many of these tokens. In the case of the *uh* token, it is reasonable that the standard deviation is low for

Table 4: Means and standard deviations for number of *uh*, *um*, and *well* tokens transcribed by WhisperX (Bain et al. 2023) and Google ASR (Google Cloud 2023).

Token	WhisperX	Google ASR
uh	**1.25** $\pm$ 2.62	0.10 $\pm$ 0.31
um	**1.65** $\pm$ 3.03	0.20 $\pm$ 0.56
well	**3.48** $\pm$ 2.76	3.51 $\pm$ 2.76

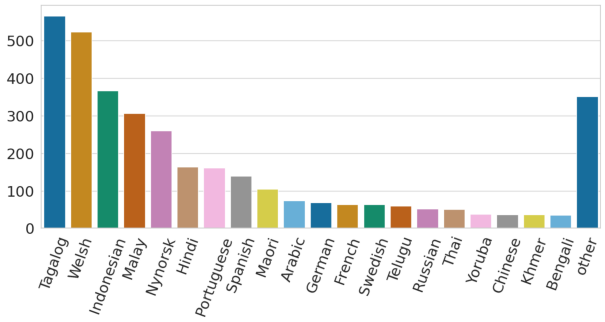

Figure 7: Distribution of the 3,521 non-English podcasts in the Spotify Podcasts Dataset (105,360, pre-filteration) (Clifton et al. 2020) as classified by WhisperX (Bain et al. 2023). We identified a total of 63 different languages present in the corpus. See section *Non-English Podcast Episodes as Identified by WhisperX* for the language makeup of the *"other"* category.

Google ASR, as there were limited *uh* tokens transcribed at all (as indicated by the mean of 0.18). The case is the same for *um*, which has a low standard deviation as it also has a low mean.

Non-English Podcast Episodes as Identified by WhisperX. The Spotify 100k dataset was designed to only contain English podcasts, according to (1) the metadata, and (2) language classification based on *langid.py*, a pre-trained multinomial naive Bayes learner (Lui and Baldwin 2011, 2012). However, we find that 3.34% of the dataset (3,521 podcasts) is comprised of non-English podcasts. We show in Figure 7 the distribution of the 3,521 non-English podcasts. We observe, upon looking at these previously misclassified podcasts, that they tend to be a blend of English and another language (as Clifton et al. (2020) also noted).

In Figure 7, of the 63 non-English languages we identified, the 42 languages shown on the figure in the long-tail *"other"* category are as follows: Yiddish, Hungarian, Dutch, Latin, Tamil, Korean, Malayalam, Urdu, Lithuanian, Javanese, Romanian, Latvian, Hebrew, Swahili, Myanmar, Vietnamese, Galician, Marathi, Afrikaans, Japanese, Norwegian, Turkish, Greek, Nepali, Shona, Finnish, Bulgarian, Sinhala, Sanskrit, Italian, Slovenian, Kannada, Breton, Punjabi, Gujarati, Haitian Creole, Hawaiian, Polish, Danish, Persian, Estonian, and Amharic.

In the LDA topics, we note a third category of language-related topics – in addition to the *content* and *discourse* categories – *language*. For example, Topic 45 contains words of multiple different languages: "tzadik" and "mitzvot" are Hebrew words, "hara" is a Hindi word, "lev" is a Bulgarian word, and "attractiveness" is an English word. We suspect that episodes which are highly weighted for this topic are primarily English – as they passed the WhisperX English language filter – interwoven with words of other languages. Examples of topics are shown in Table 5

The Importance of High-Quality Transcription. The podcasts were originally transcribed using Google Automatic Speech Recognition (ASR) (Google Cloud 2023; Clifton et al. 2020). Per Clifton et al. (2020), Google ASR had a sample word error rate (WER) of 18.1% on the podcasts. Comparably, WhisperX shows a WER of 13.8% on the Switchboard dataset (Linguistic Data Consortium 2002), which is also a conversational dataset. Google ASR tends to transcribe less disfluent discourse tokens – such as *uh*, *um*, and *well* – than WhisperX, as shown in Table 4.[9]

These disfluencies, as well as improved overall transcription quality, support more fine-grained analysis of how people communicate. Indeed, "[s]pontaneous human speech is notoriously disfluent" (Brennan and Schober 2001). While some podcasts are scripted, and thus match the traditional training data of ASR systems – audiobooks – (Panayotov et al. 2015), many are not, and are instead unscripted, and consist of spontaneous, conversational, disfluent speech (e.g., interviews, talk shows, etc.). People "are highly sensitive to hesitation disfluencies in speech" (Corley and Stewart 2008), as "words preceded by disfluency [are] more likely to be remembered" by listeners (Corley, MacGregor, and Donaldson 2007). The different types of disfluency even produce different levels of *memory boosts* (Diachek and Brown-Schmidt 2023). This is important, as disfluencies are common, occurring at a rate of approx. 4-6 disfluent words per 100 words (Tree 1995; Branigan, Lickley, and McKelvie 1999). On the speaker side, "disfluencies are associated with an increase in planning difficulty" (Bortfeld et al. 2001). "[S]peakers use uh and um to announce that they are initiating what they expect to be a minor (uh), or major (um), delay in speaking" (Clark and Tree 2002), and this makes sense, as "[disfluencies] precede relatively unpredictable lexical items [and] relatively infrequent lexical items" (Beattie and Butterworth 1979). Hence, towards characterizing the podcast content, we aim to retain disfluencies in our audio transcriptions using WhisperX.

Disfluencies are also interesting in the context of gender, where "filled pauses may serve to 'hold the floor"' (Shriberg 1996) – i.e., to verbally take up more time in conversation. Additionally, there is a difference in disfluent, filler speech production with respect to gender (Bortfeld et al. 2001). The podcasts, being spoken content and representing human conversational speech, thus provide a new opportunity to study gendered speech differences.

Large-Scale Confirmation of Small-Scale Studies

OTHER Module Details. *Duration* refers to the floating-point time in minutes per episode. This feature was provided as part of the Spotify Podcasts Dataset (Clifton et al. 2020). *Speech Rate* is approximated by measuring the integer number of words in each 10-minute truncated WhisperX transcript. A transcript with a higher word count has more words spoken in the same amount of time (10 minutes) as a transcript with a lower word count.

CONVERSATIONAL PARSER Module Details. We use a state-of-the-art parsing model, *english-fisher-*

[9]Clifton et al. (2020) notes: *"We also anticipate that the state of the art in automatic speech recognition will improve in the coming years, allowing for more accurate automatic transcriptions."*

Table 5: Topic modeling with LDA: a few of the 100 topics and the top 10 weighted words for that topic.

Topic Number	Category	Subcategory	Top 10 Words for Topic
Topic 5	Content	Crime	police, crime, murder, case, killer, serial, crimes, criminal, victim, killed
Topic 20	Content	Football	jones, bowl, dallas, austin, smith, nfl, cowboys, giants, miami, eagles
Topic 22	Content	Food	coffee, drink, drinking, wine, party, tea, bar, chocolate, glass, cheese
Topic 34	Content	Medical	patients, pain, patient, disease, treatment, injury, risk, test, type, symptoms
Topic 57	Content	Church	music, song, church, songs, album, art, mary, band, love, bible
Topic 66	Content	History	war, military, army, oil, elizabeth, russian, ii, soldiers, edward, russia
Topic 70	Content	Cars	car, drive, cars, driving, road, truck, tesla, train, traffic, miles
Topic 85	Content	Diet	food, eat, eating, weight, body, fat, day, diet, healthy, nutrition
Topic 54	Discourse	Informal	get, like, know, right, people, going, podcast, make, want, one
Topic 60	Discourse	Informal	going, know, think, get, got, one, really, good, well, yeah
Topic 62	Discourse	Informal	like, know, really, going, people, want, think, get, things, life
Topic 88	Discourse	Formal	one, said, would, man, see, way, says, let, say, us
Topic 100	Discourse	Informal	like, yeah, know, oh, right, got, okay, think, one, get
Topic 45	Language	–	tzadik, supercross, hara, mitzvot, midas, lev, tomek, barsha, attractiveness, marv

Table 6: Code and licenses.

	Link	License
The Spotify Podcast Dataset	https://podcastsdataset.byspotify.com/	Creative Commons Attribution 4.0 International License
WhisperX	https://github.com/m-bain/whisperX	BSD-4-Clause License
CountVectorizer	https://scikit-learn.org/stable/modules/generated/sklearn.feature_extraction.text.CountVectorizer.html	BSD License
LatentDirichletAllocation	https://scikit-learn.org/stable/modules/generated/sklearn.decomposition.LatentDirichletAllocation.html	BSD License
NLTK	https://www.nltk.org/howto/corpus.html?highlight=stopwords	Apache License 2.0
inaSpeechSegmenter	https://github.com/ina-foss/inaSpeechSegmenter	The MIT License
english-fisher-annotator	https://github.com/pariajm/english-fisher-annotations	None

Table 7: Significant correlations between duration, speech rate, and gender.

	Women	Men
Duration	-0.17	0.12
Speech Rate	–	0.15

annotator Jamshid Lou and Johnson (2020), for parsing sentences and obtaining part-of-speech (POS) counts for each 10-minute transcript. As we used WhisperX for the transcriptions, we have high-quality transcripts – which also include more disfluencies (i.e., *um*, *uh*, and more), see Figure 4 – and therefore use this parsing model, which was designed for use in disfluent, conversational settings.

The parsing model, *english-fisher-annotator*, specializes in handling annotation of the edited part-of-speech nodes, which often arises in conversational, spontaneous speech. The model was evaluated on the Switchboard dataset (Godfrey and Holliman 1997; Mitchell et al. 1999), and is scored based on its performance on the disfluent edited, interjection, and parenthetical node types. In the sentence fragment *It was cold, oh I think, it was hot outside...*, *It was cold* is an edited node, *oh* is an interjection, and *I think* is a parenthetical; deleting these from the sentence would form a fluent sentence, hence these are the disfluent node types. The model scores $P = 92.5$, $R = 97.2$, and $F = 94.8$ for edited, interjection, and parenthetical nodes (Jamshid Lou and Johnson 2020).

We annotated all of the truncated transcripts to obtain their parse trees on a sentence-level. We ran *english-fisher-annotator* on a single machine with 2 NVIDIA TITAN Xp GPUs. We obtain the counts for each POS by counting the number of times that label occurs across all the parse trees for each 10-minute transcript. We obtain the parse trees on a sentence level, and truncate all sentences in each transcript to 300 tokens for compatibility with *english-fisher-annotator*. We use a subset of all possible POS labels (see Figure 10) for our analysis (Taylor, Marcus, and Santorini 2003; Jamshid Lou and Johnson 2020).

Taylor, Marcus, and Santorini (2003) describe the creation of the Penn Treebank-3 dataset for evaluating the models. Charniak and Johnson (2001) formalize the evaluation metrics for the parsing-based disfluency annotation task. Disfluency parsing is an established line of work (Johnson and Charniak 2004; Honnibal and Johnson 2014; Tran et al. 2017; Lou, Wang, and Johnson 2019; Jamshid Lou and Johnson 2020).

As shown in Equation 5, the model classifies all of the spans in a string, from position i to position j with a label l based on the classification scores for each span. The model then calculates the score for each parse tree, $s(T)$, by summing $s(i, j, l)$.

$$s(T) = \sum_{(i,j,l) \in T} s(i, j, l) \tag{5}$$

Then, out of all the possible parse trees, $s(T)$, the highest-scoring parse tree, $\hat{T}$, is selected as the parse tree for that sentence, as shown in Equation 6.

$$\hat{T} = \underset{T}{\operatorname{argmax}}\ s(T) \tag{6}$$

We use the *swbd_fisher_bert_Edev.0.9078* model checkpoint.

Table 8: Significant correlations between content topics and content topics.

Topic N	Topic M	r	Topic N Word List	Topic N Categories	Topic M Word List	Topic M Categories
Topic 3	Topic 34	0.10	women, woman, men, baby, pregnant, girls, men, doctor, health, birth	Content - Pregnancy	patients, pain, patient, disease, treatment, injury, risk, test, type, symptoms	Content - Medical
Topic 5	Topic 53	0.31	police, crime, murder, case, killer, serial, crimes, criminal, victim, killed	Content - Crime	would, family, years, could, children, father, life, time, home, young	Content - Family
Topic 11	Topic 63	0.34	data, new, technology, public, bill, theory, science, system, security, article	Content - Technology/ Politics	people, world, black, country, america, states, history, white, american, united	Content - USA
	Topic 66	0.12			war, military, army, oil, elizabeth, russian, ii, soldiers, edward, russia	Content - European History
Topic 63	Topic 66	0.26	people, world, black, country, america, states, history, white, american, united	Content - USA	war, military, army, oil, elizabeth, russian, ii, soldiers, edward, russia	Content - European History
Topic 12	Topic 72	0.14	business, money, company, market, buy, right, million, companies, pay, sell	Content - Business	bitcoin, adam, people, crypto, show, coin, network, mining, coins, meister	Content - Cryptocurrency
Topic 13	Topic 80	0.33	club, soccer, la, phil, well, mexico, de, real, lucas, madrid	Content - Soccer	goal, goals, league, season, cup, yeah, points, obviously, chelsea, premier	Content - Soccer
Topic 49	Topic 13	0.13	game, know, think, team, going, mean, play, year, one, good	Content - Sports	club, soccer, la, phil, well, mexico, de, real, lucas, madrid	Content - Soccer
	Topic 20	0.41			jones, bowl, dallas, austin, smith, nfl, cowboys, giants, miami, eagles	Content - Football
	Topic 25	0.30			state, florida, college, texas, south, north, carolina, michigan, georgia, ohio	Content - USA States
Topic 22	Topic 71	0.13	coffee, drink, drinking, wine, party, tea, bar, chocolate, glass, cheese	Content - Food/Drink	christmas, sex, girl, hair, love, get, date, girls, let, wear	Content - Dating

Table 9: Significant correlations between POS and gender.

Part-of-Speech	Women	Men
Edited	-0.14	0.16
Parenthetical	-0.11	0.12
Adjective Phrase	0.13	–
Noun Phrase	–	0.17
Prepositional Phrase	-0.1	0.14

Table 10: Significant correlations between parts-of-speech, and formal and informal discourse (as determined by their top weighted words).

Part-of-Speech	Topic 100: Discourse - Informal	Topic 88: Discourse - Formal
Interjection	0.86	-0.20
Edited	0.37	-0.22
Parenthetical	0.17	-0.22
Adjective Phrase	0.17	-0.19
Adverb Phrase	0.22	-0.30
Noun Phrase	0.14	-0.12
Prepositional Phrase	-0.43	-
Simple Declarative Clause	0.19	-0.18
Verb Phrase	0.17	-0.18

How are gender, duration, and speech rate related? Starting with the first row in Table 7, we see that *Duration* and *women* have a negative correlation (-0.17), and *Duration* and *men* have a positive correlation (0.12). This indicates that the **more** minutes in duration a podcast episode is, there tends to be, then, **more** seconds of men speech in the first 30 seconds of the podcast episode. Conversely, for the case of *Duration* and the feature *Gender - women*, the correlation is -0.17, indicating that the **more** minutes in duration the podcast is, there tends to be **less** seconds of women speech in the first 30 seconds of the podcast episode.

In the second row, we see that *Speech Rate* and *women* do not have a significant correlation, whilst *Speech Rate* and *men* have a significant positive correlation (0.15). This indicates that the more masculine a podcast is, the faster that the rate of speech is in that podcast.

These findings are consistent with Leaper and Ayres (2007) and James and Drakich (1993), who also find that men are overall more "talkative" than women, which relates to men "hold[ing] the floor" (Shriberg 1996). Shriberg (1996) notes that this may be due to gender being confounded with other variables – such as education level and occupation – that men are able to "hold the floor" longer than women.

Which topics are related? In Table 8, we can see that similar topics have positive correlations with each other. Topics 3 (Pregnancy) and 34 (Medical) are positively correlated. Topics 5 (Crime) and 53 (Family) are correlated. Topics 11 (Technology/Politics), 63 (USA), and 66 (European History) are all positively correlated with each other. Topics 12 (Business) and 72 (Cryptocurrency) are positively correlated. Topic 49 (Sports) is positively correlated with Topics 13 (Soccer), 20 (Football), and 25 (USA States). Topics 22 (Food/Drink) and 71 (Dating) are positively correlated. These correlations imply that there may be some clustering structure to the topics, and deserves further study.

How do parts-of-speech vary by gender? In Table 9, disfluent parts-of-speech (edited and parenthetical (Jamshid Lou and Johnson 2020)) are negatively correlated with *women*, and positively correlated with *men*, indicating that men tend to be more disfluent in their speech. *Adjective Phrases* are positively correlated with *women*, while *Noun Phrases* and *Prepositional Phrases* are positively correlated with *men*. This indicates that women tend to use more descriptive words (adjective phrases) in their speech. Shriberg (1996) observed an association between increased filled pauses and men, and states that "filled pauses may serve to 'hold the floor,'" but also that gender is confounded with other variables such as education level and occupation. Bortfeld et al. (2001) also found that men pro-

Table 11: Significant correlations between discourse topics. Gender correlation labels for Topics N and M are assigned based on significant correlations from Table 1.

Topic N	Topic M	r	Topic N Word List	Topic N Categories	Topic M Word List	Topic M Categories
Topic 60	Topic 62	-0.34	going, know, think, get, got, one, really, good, well, yeah	Discourse - Informal (Men)	like, know, really, going, people, want, think, get, things, life	Discourse - Informal (Women)
Topic 100	Topic 60	-0.13	like, yeah, know, oh, right, got, okay, think, one, get	Discourse - Informal	going, know, think, get, got, one, really, good, well, yeah	Discourse - Informal (Men)
	Topic 62	-0.29			like, know, really, going, people, want, think, get, things, life	Discourse - Informal (Women)
	Topic 88	-0.12			one, said, would, man, see, way, says, let, say, us	Discourse - Formal

duced more fillers than women in their speech.

How do parts-of-speech vary for informal and formal discourse? In Table 10, we see that *Topic 100*, informal language, tends to have more disfluencies (*Interjection*, *Edited*, and *Parenthetical* parts-of-speech (Jamshid Lou and Johnson 2020)) than *Topic 88*, formal language, as shown in Table 1.

Topic 100, informal language, is highly correlated (0.86) with increased *Adjective Phrases*, *Adverb Phrases*, and *Noun Phrases*, while *Topic 88*, formal language, is *not* correlated with these parts-of-speech. *Topic 100*, informal language, is also negatively correlated with *Prepositional Phrases*. This means that informal language tends to be more descriptive, as characterized by more adjective phrases, adverb phrases, and noun phrases.

Are discussion styles distinct? Table 11 examines the relationship between informal and formal discourse topics. Starting with the first row, the correlation between *Topic 60* and *Topic 62* is -0.34. *Topic 62* is a topic which is characterized by informal speech, as shown by its word list, and from Table 1, *Topic 62* has a positive correlation with *women* (0.33) and a negative relationship with *men* (-0.28), making it a *women* topic. This contrasts with *Topic 60*, which is primarily a *men* topic. Hence, the data indicates that men informal language and women informal language tend not to co-occur. This aligns with the correlation value for the *men* and *women* features: -0.76. *Topic 100* is extremely informal, as it includes swear words. We see that it is distinct from the other informal discourse topics, *Topic 60* and *Topic 62*, as it has -0.13 and -0.29 correlation values with these topics.

Impact of Lemmatization on LDA with Non-Contextual Embeddings

We find that lemmatization before the creation of the non-contextual bag-of-words embeddings does not have a significant impact on the quality of the discourse topics created by LDA. Looking to Tables 12 and 13, we see that discourse topics are still formed with and without the lemmatization step, and that these topics have significant correlations with women and men.

While the topics formed with lemmatization are different than the topics formed without lemmatization, we still see patterns of gendered speech that emerge within these discourse topics, as indicated by r. Even on a smaller sample size of 10,000 podcasts, many of the correlations are high ($|r| > 0.20$),

WEAT versus SEAT

WEAT (Caliskan, Bryson, and Narayanan 2017) words are advantageous as compared to SEAT (May et al. 2019) sentences, because "the context is artificial, which does not reflect the natural usage of a word." (Nadeem, Bethke, and Reddy 2020).

D-WEAT with BERTopic Discourse Topics

We run the D-WEAT experiment with the discourse topics formed via BERTopic with contextual embeddings (BERT, ChatGPT, Llama). **Target Words [T_w, T_m].** We form T_w and T_m in the same way, using our discourse topics from BERTopic (see Table 2):

- $T_w = \{$*life, know, things, really, people, feel, want, love, way, person*$\}$ These are the top weighted words *post-filtering* from Topic 2,[10] which is significantly positively correlated with *women* and negatively correlated with *men*, representing the feminine discourse style.
- $T_m = \{$*like, yeah, oh, right, podcast, got, going, think, okay, f***ing*$\}$ We censor the last word for presentation, but not in the experiment. These are the top 10 weighted words *post-filtering* from Topic 0, which is significantly positively correlated with *men* and negatively correlated with *women*, representing the masculine discourse style.

LLM Representation. We use `text-embedding-3-large` from Open AI via the API with a single call. **Impact of T_w and T_m.** We study the impact of varying T_w and T_m formed via BERTopic with contextual embeddings (BERT, ChatGPT, Llama). We fix $\tau = 30.0$ and $\gamma = 6$. We make 1 API call. We use 1 seed. We find that for S_w, the man percentage is 100% and the woman percentage is 0%. For S_m, the man percentage is approx. 22% and the woman percentage is approx. 78%. This finding indicates that the gendered discourse words discovered via BERTopic are represented in a gender-imbalanced way in the embedding model, and hence, this is a masculine default.

D-WEAT with Llama Embeddings

We run the D-WEAT experiment with Llama embeddings for the LLM representation. We obtain the embeddings from `Llama-3.1-8B-Instruct` using the PromptEOL method (Jiang et al. 2023). **Impact of Embeddings.** We study the impact of varying the embedding model. We fix

[10] Topic 5 is only positively correlated with *women*, and has no significant correlation with *women*; thus, we use Topic 0, as it has significant correlations for both *women* $(+)$ and *men* $(-)$.

Table 12: **LDA with Non-Contextual Embeddings (Bag-Of-Words) *with Lemmatization***: Significant correlations between gender features and topic features for *discourse topics only* (content topics are omitted).

Topic N	Gender	r	Topic N Word List	Topic N Categories	Topic N Gender
Topic 30	Women	0.17	like, know, yeah, go, get, think, really, say, want, right	Discourse	Women
	Men	-0.12			
Topic 75	Women	-0.31	get, go, know, think, yeah, game, year, play, good, one	Discourse	Men
	Men	0.25			
Topic 91	Women	0.13	people, go, thing, really, know, get, want, think, work, time	Discourse	Women
	Men	-0.11			

Table 13: **LDA with Non-Contextual Embeddings (Bag-Of-Words) *without Lemmatization***: Significant correlations between gender features and topic features for *discourse topics only* (content topics are omitted).

Topic N	Gender	r	Topic N Word List	Topic N Categories	Topic N Gender
Topic 45	Women	0.29	know, like, really, people, going, think, want, things, get, kind	Discourse	Women
	Men	-0.23			
Topic 74	Women	-0.25	think, know, going, game, got, team, yeah, good, year, one	Discourse	Men
	Men	0.20			
Topic 95	Women	-0.12	one, going, well, get, got, would, time, yeah, back, go	Discourse	Men
	Men	0.06			

$\tau = 30.0$ and $\gamma = 6$. We make 1 API call. We use 1 seed. We find that for S_w, the man percentage is 70% and the woman percentage is 30%. For S_m, the man percentage is 50% and the woman percentage is 50%. This finding indicates that the gendered discourse words discovered via BERTopic are represented in a gender-imbalanced way in the embedding model – in that men obtain a stable representation (no gap between women and man percentages) while women do not (larger gap between women and man percentages). Hence, this is a masculine default.