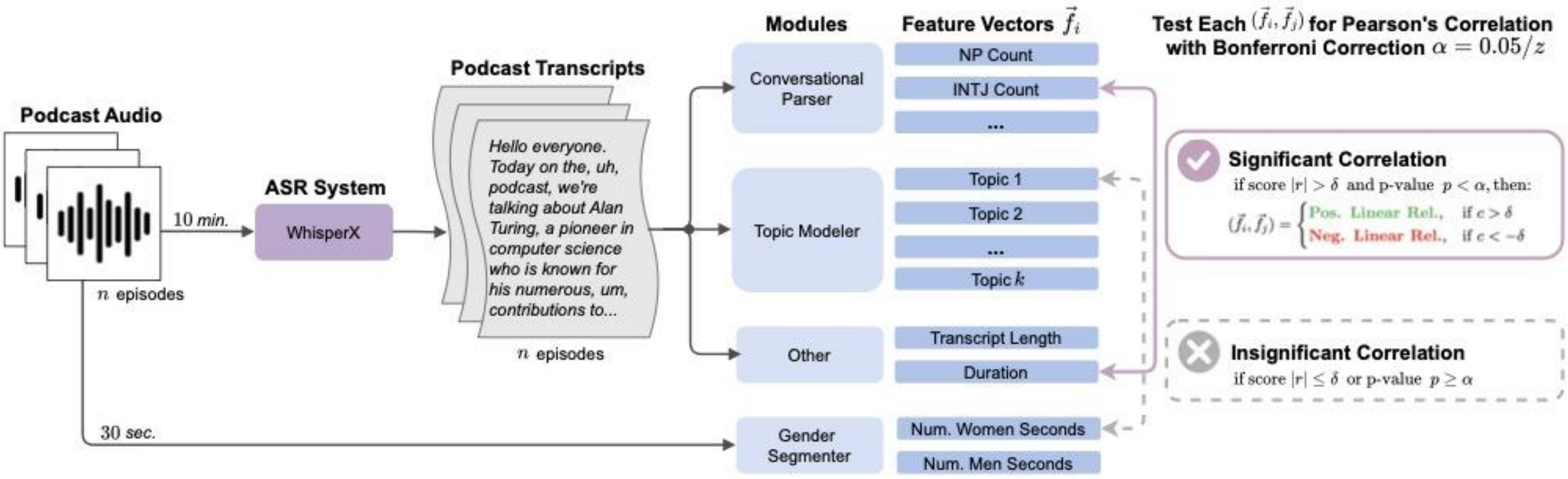
Masculine Defaults via Gendered Discourse in Podcasts and Large Language Models

Maria Teleki, Xiangjue Dong, Haoran Liu, James Caverlee

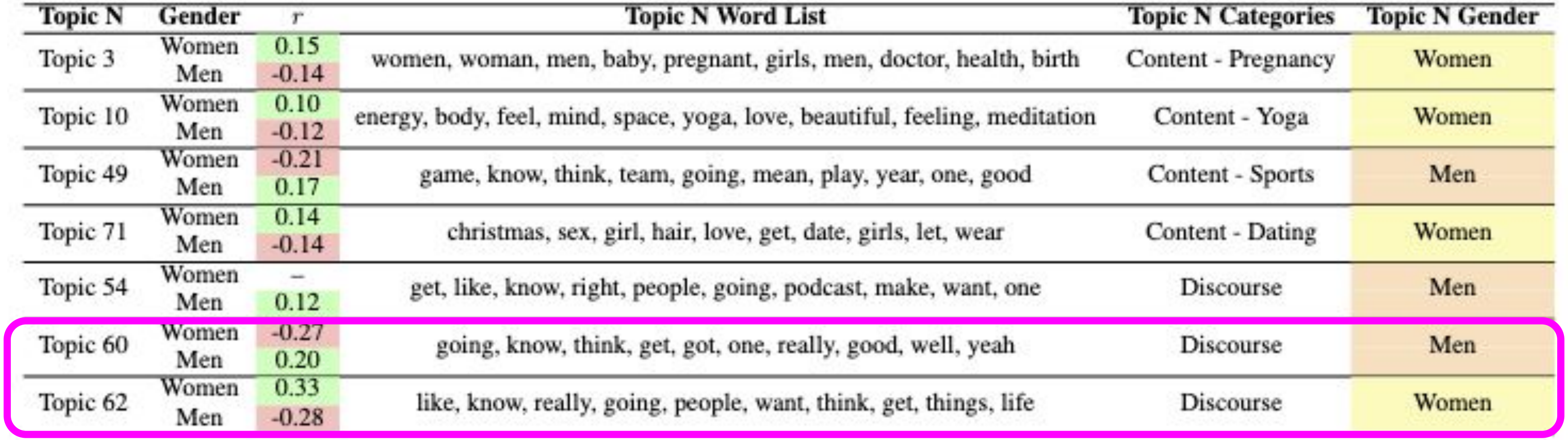
Texas A&M University
{mariateleki, xj.dong, liuhr99, caverlee}@tamu.edu

Gendered Discourse Correlation Framework (GDCF)


Podcast Audio
n episodes
10 *min.*
30 *sec.*
ASR System
WhisperX
Podcast Transcripts
Hello everyone. Today on the, uh, podcast, we're talking about Alan Turing, a pioneer in computer science who is known for his numerous, um, contributions to...
n episodes
Modules
Conversational Parser
Topic Modeler
Other
Gender Segmenter
Feature Vectors $\vec{f}_i$
NP Count
INTJ Count
...
Topic 1
Topic 2
...
Topic k
Transcript Length
Duration
Num. Women Seconds
Num. Men Seconds
Test Each $(\vec{f}_i, \vec{f}_j)$ for Pearson's Correlation with Bonferroni Correction $\alpha = 0.05/z$
Significant Correlation
if score $|r| > \delta$ and p-value $p < \alpha$, then:
$(\vec{f}_i, \vec{f}_j) = \begin{cases} \text{Pos. Linear Rel.,} & \text{if } c > \delta \\ \text{Neg. Linear Rel.,} & \text{if } c < -\delta \end{cases}$
Insignificant Correlation
if score $|r| \leq \delta$ or p-value $p \geq \alpha$

We obtain ***audio and text-based features*** for the ***Spotify Podcasts***, and test for **significant correlations between these features**.

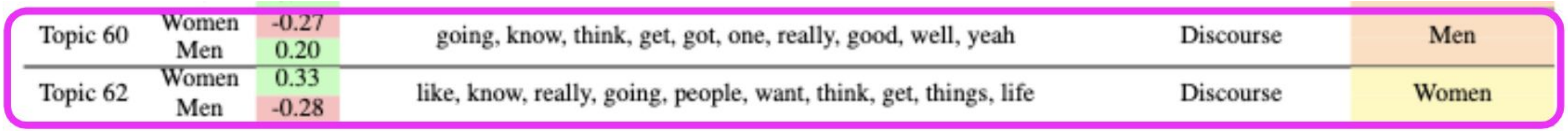
Gendered Discourse Correlation Framework (GDCF)



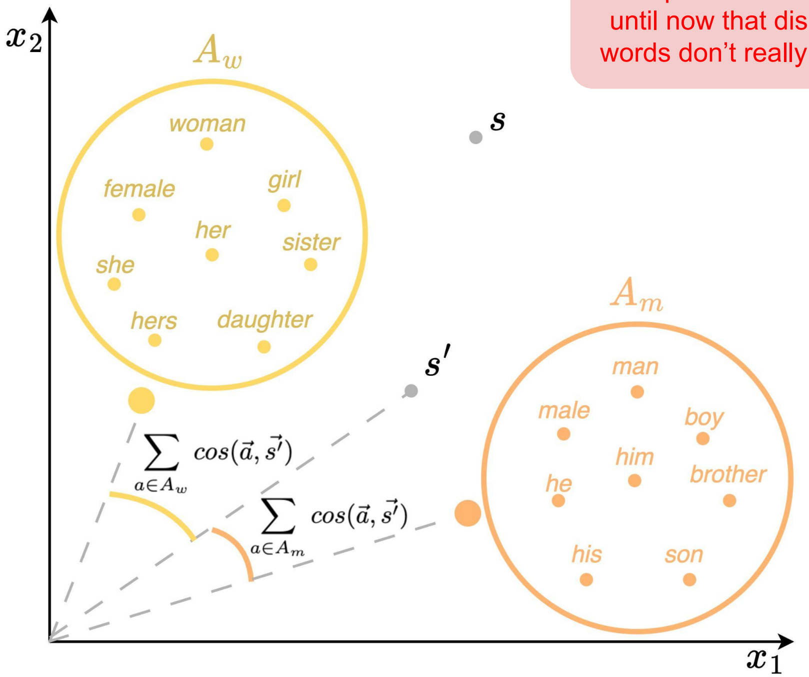
Topic N	Gender	r	Topic N Word List	Topic N Categories	Topic N Gender
Topic 3	Women	0.15	women, woman, men, baby, pregnant, girls, men, doctor, health, birth	Content - Pregnancy	Women
	Men	-0.14			
Topic 10	Women	0.10	energy, body, feel, mind, space, yoga, love, beautiful, feeling, meditation	Content - 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			

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

Discourse Word-Embedding Association Test (D-WEAT)

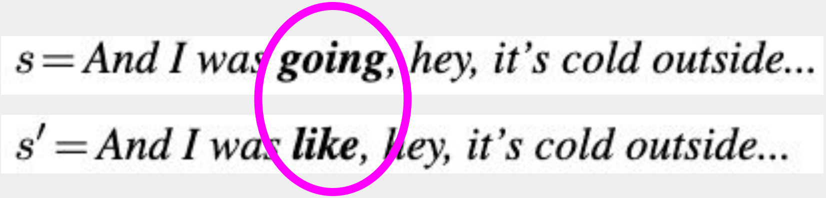


Topic 60	Women	-0.27	going, know, think, get, got, one, really, good, well, yeah	Discourse	Men
	Men	0.20			
Topic 62	Women	0.33	like, know, really, going, people, want, think, get, things, life	Discourse	Women
	Men	-0.28			



People have assumed until now that discourse words don't really matter!

We set up an experiment to measure: What happens if we swap the discourse words? Does the sentence "move closer" to the other gender?



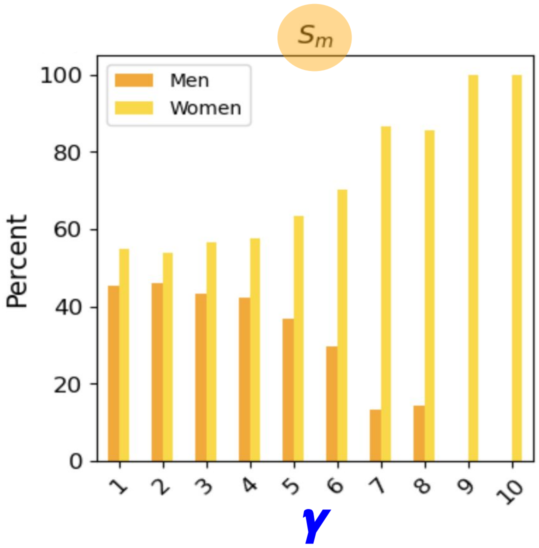
$s =$ *And I was* ***going****, hey, it's cold outside...*

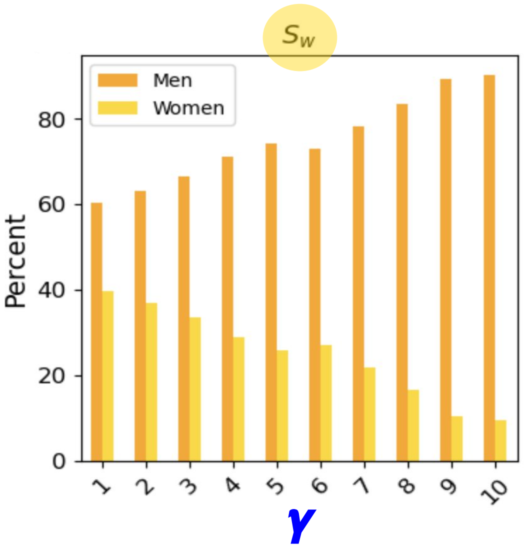
$s' =$ *And I was* ***like****, hey, it's cold outside...*

S_m: masculine → feminine discourse word replacement

S_w: feminine → masculine discourse word replacement

Impact of γ



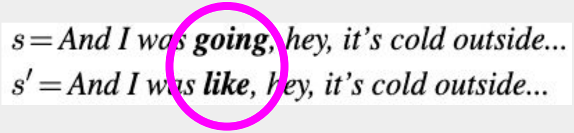


We see that the embedding moves towards the **feminine concept** in the embedding space.

We see that the embedding moves towards the **masculine concept** in the embedding space.

We also see that the overall ***gap is bigger for the S_w sentences than the S_m sentences*** – meaning **men have a more robust discourse embedding representation than women.**

What is γ?

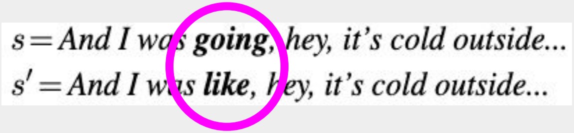


$s =$ *And I was* ***going,*** *hey, it's cold outside...*

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

In this example, γ=1, because we do 1 discourse word replacement.

What are S_m and S_w?



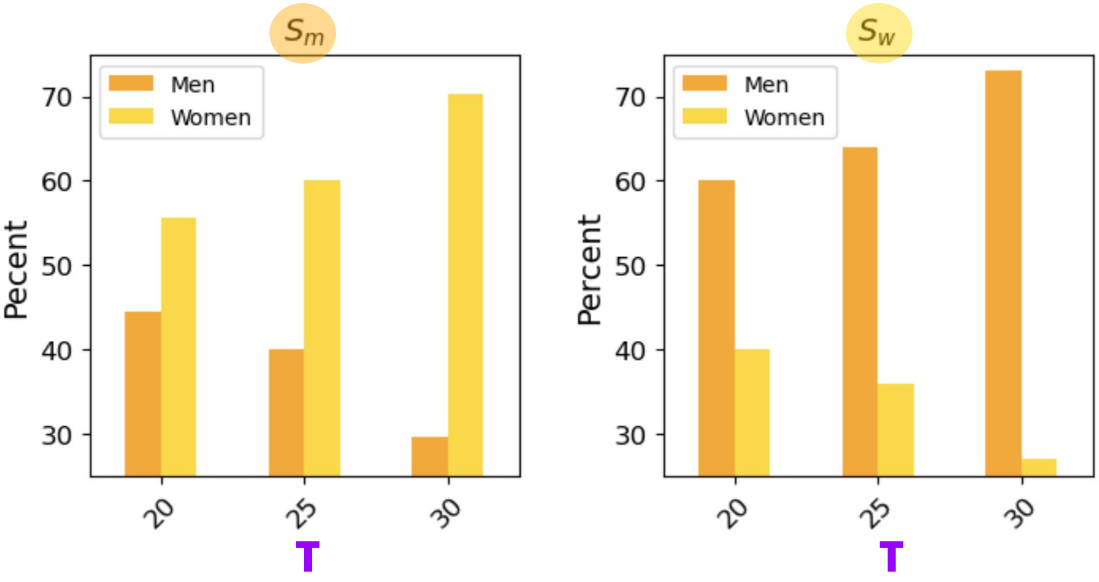
$s =$ *And I was* ***going,*** *hey, it's cold outside...*

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

S_m: masculine → feminine discourse word replacement

S_w: feminine → masculine discourse word replacement

Impact of τ

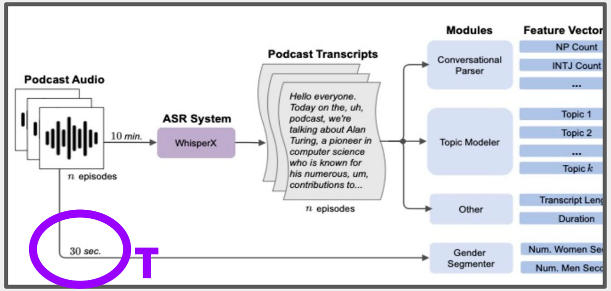


We see that the embedding moves towards the **feminine concept** in the embedding space.

We see that the embedding moves towards the **masculine concept** in the embedding space.

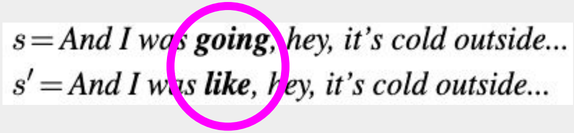
We also see that the overall ***gap is bigger for the*** S_w ***sentences than the*** S_m ***sentences*** – meaning **men have a more robust discourse embedding representation than women.**

What is τ?



This variable is τ, the # of seconds we take from the podcast audio for our gender features in the GDCF pipeline.

What are S_m and S_w?



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

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

S_m: masculine → feminine discourse word replacement

S_w: feminine → masculine discourse word replacement

Why does it matter that men have a more robust discourse embedding representation than women?

Men can get better performance on LLM tasks (Cao et al. 2022; Kaneko and Bollegala 2021) – i.e. men have **better access to information**.

This fact is a **representational harm** (Blodgett et al. 2020). Also, this knowledge advances our understanding of the **current hegemonic masculine strategy** (Connell 1995, 1987) and the **current technomasculine strategy** (Cooper 2000; Lockhart 2015; Bulut 2020) in the technology domain.

D-WEAT joins 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) **as an intrinsic metric that can be used to regulate bias in LLMs.**