דוח סופי – אבולוציית השירים

דאטה- קישור לGitHub, קישור לGenius.com

משכנו נתונים מתוך אתר genius, שהיא פלטפורמה למציאת lyrics לשירים בלעז. (מעין "שירונט")

ביצענו זאת על ידי API של האתר המאפשר למשוך נתונים בסיסיים כגון: שם השיר, אומן, שנת הוצאה, מילות השיר עצמו וכו...

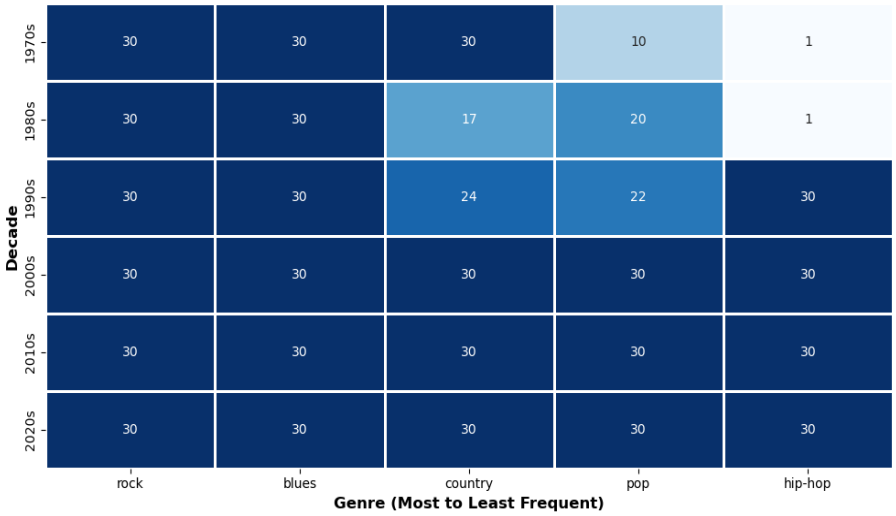
החלטנו למשוך דאטה של שירים לפי חלוקה לקבוצות של 5 ז'אנרים נבחרים בלבד ולפי עשורים בין 1970 ל2020 כך שכל קבוצה תמנה 30 שירים שונים.

משיכת השירים מהAPI מבוצעת לפי סדר הפופולריות באתר (כך שהשירים בדאטה מייצגים שירים פופולריים יחסית ולא נידחים)

לאחר משיכת הדאטה הגולמי, יצרנו משתנים נוספים המבוססים על הנתונים הגולמיים כגון: מילות השיר בצורה נקייה (ללא הערות וטקסט של מבנה השיר), יחסי מילים פוגעניות ומילות תואר ביחס לכלל המילים, משתנים הקשורים למבנה השיר ועוד ועוד. (ניתן לראות ב dataset data dictionary)

ומתוך כלל המשתנים החדשים יצרנו את הדאטה סט שאיתו עבדנו לאורך הפרויקט.

Figure 0: **Song Count by Decade and Top Genres (Ordered by Volume)**



סך הכל 785 שירים בדאטה-סט. ניתן לשים לב שישנם קבוצות בעלות מדגם קטן מאוד ולכן בניתוח חלק מהנתונים ביצענו ניקוי של קבוצות אלה על מנת שלא ישפיעו על התוצאות.

שאלת המחקר

השאלה המרכזית אליה נתייחס בדו"ח זה היא: **"האם השירים הפופולריים שאנחנו שומעים הולכים ונעשים דלים ופשוטים יותר מבחינה מילולית ומבנית לעומת שירים מעשורים קודמים?"**

נענה על שאלה זו בעזרת מספר שאלות ביניים בהן ניעזר בגרפים בכדי לנסות ולהגיע לתשובה לשאלה זו:

שאלה 1 - האם יש שינוי באורך השיר בעשורים השונים? האם זה חוזר על עצמו בכל הז'אנרים?

Figure 2:

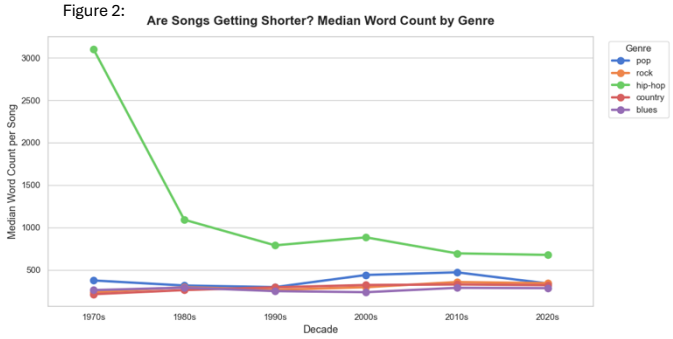


Figure 1:

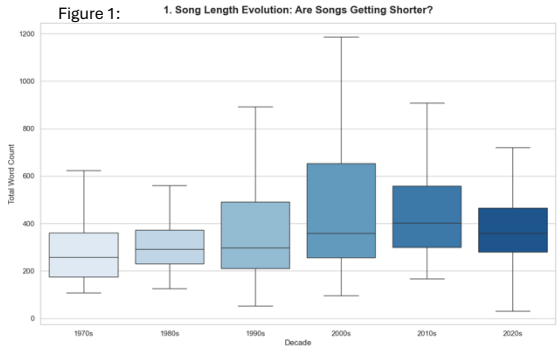
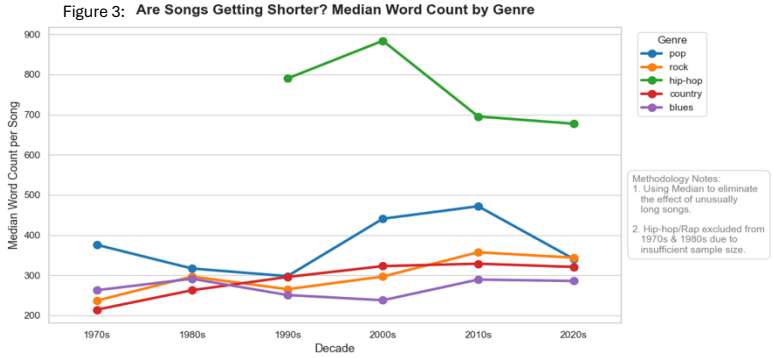


Figure 3:



גרף 1 – Box plot של כמות המילים לשיר לפי עשורים , גרף 2 – ערכים חציוניים של כמות המילים לפי ז׳אנרים לכל עשור (כל הדאטה) , גרף 3 – כמו גרף 2 לאחר ניקוי מדגמים קטנים.

שאלה 2 – האם ניתן לראות שינוי בעושר המילולי בעשורים השונים?

Figure 4:

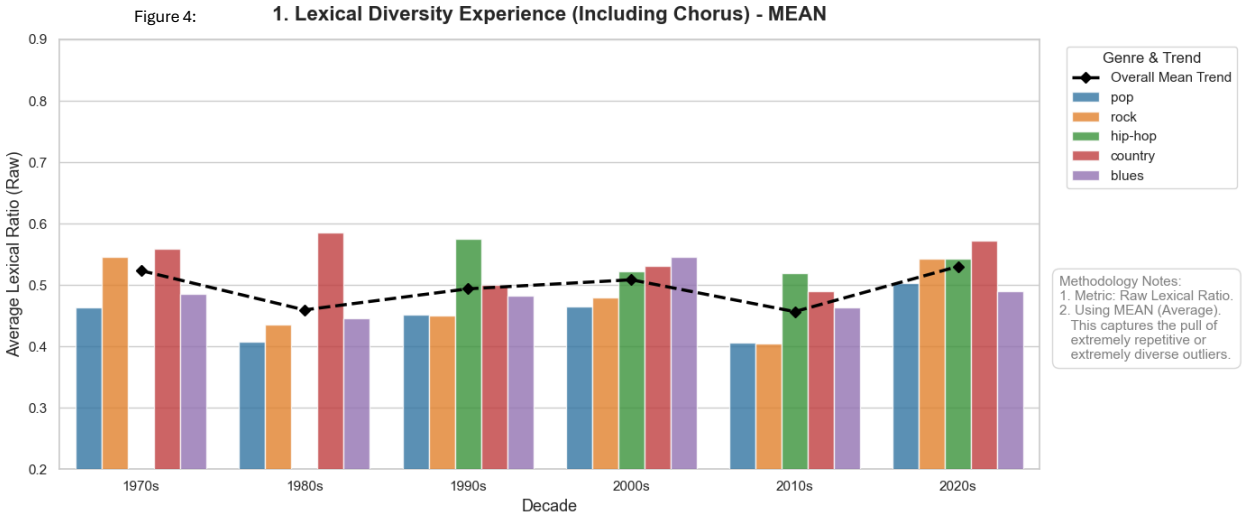
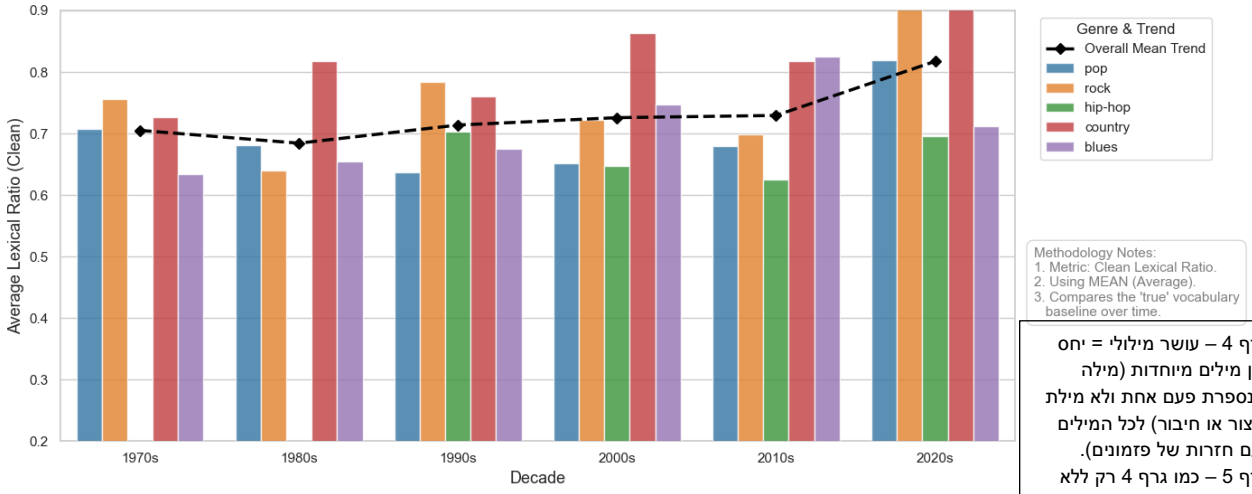


Figure 5: **2. True Lexical Diversity (Excluding Chorus) - MEAN**

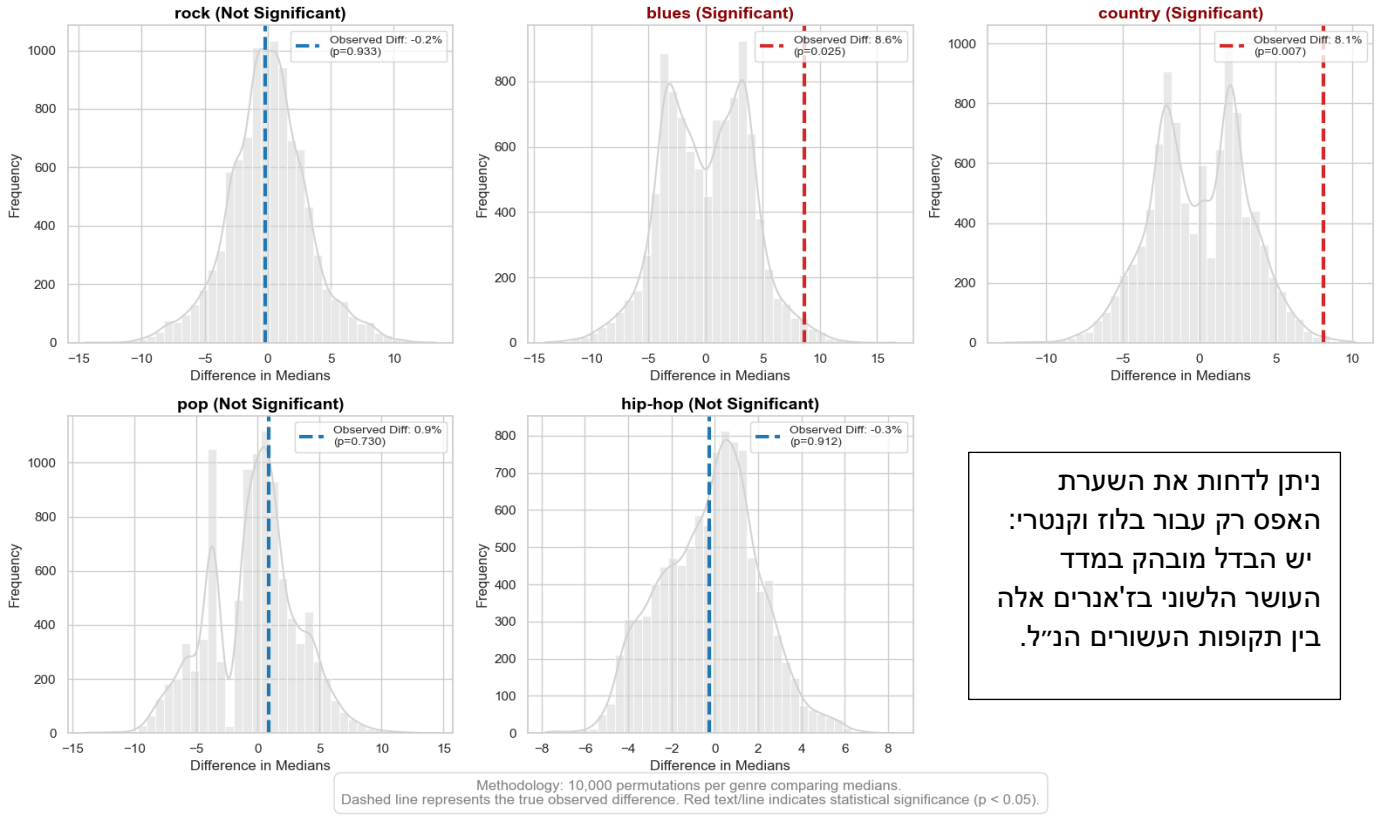


גרף 4 – עושר מילולי = יחס בין מילים מיוחדות (מילה שנספרת פעם אחת ולא מילת עיצור או חיבור) לכל המילים (עם חזרות של פזמונים).
גרף 5 – כמו גרף 4 רק ללא חזרות של פזמונים.

ניתוח הגרפים מעלה תמונה מפתיעה: בעוד שבמדד העושר הלשוני הכולל את חזרות הפזמונים לא נראית מגמה ברורה לאורך השנים, בגרף המנטרל את חזרות הפזמון מתגלה **מגמת עלייה בעושר הלשוני** .המסקנה היא שבניגוד לסטיגמה הרווחת, מגוון המילים הייחודיות (האוצר המילולי הנקי) שהאומנים משלבים בשירים דווקא הולך וגדל עם העשורים.

רצינו לבדוק את המובהקות הסטטיסטית של התוצאות, במבחן פרמוטציות למדד העושר הלשוני לפי שתי קבוצות שנים (1970-1990 ו 2000 -2020), בבחירת MOE של הפרש חציוני הקבוצות לא היה ניתן לדחות את השערת האפס (אין הבדל בין הקבוצות ביחס לעושר לשוני).
אך בניתוח של כל ז׳אנר בנפרד מתקבלות התוצאות הבאות:

Figure 6: **Permutation Tests by Genre: Median Lexical Richness (New vs. Old Era)**

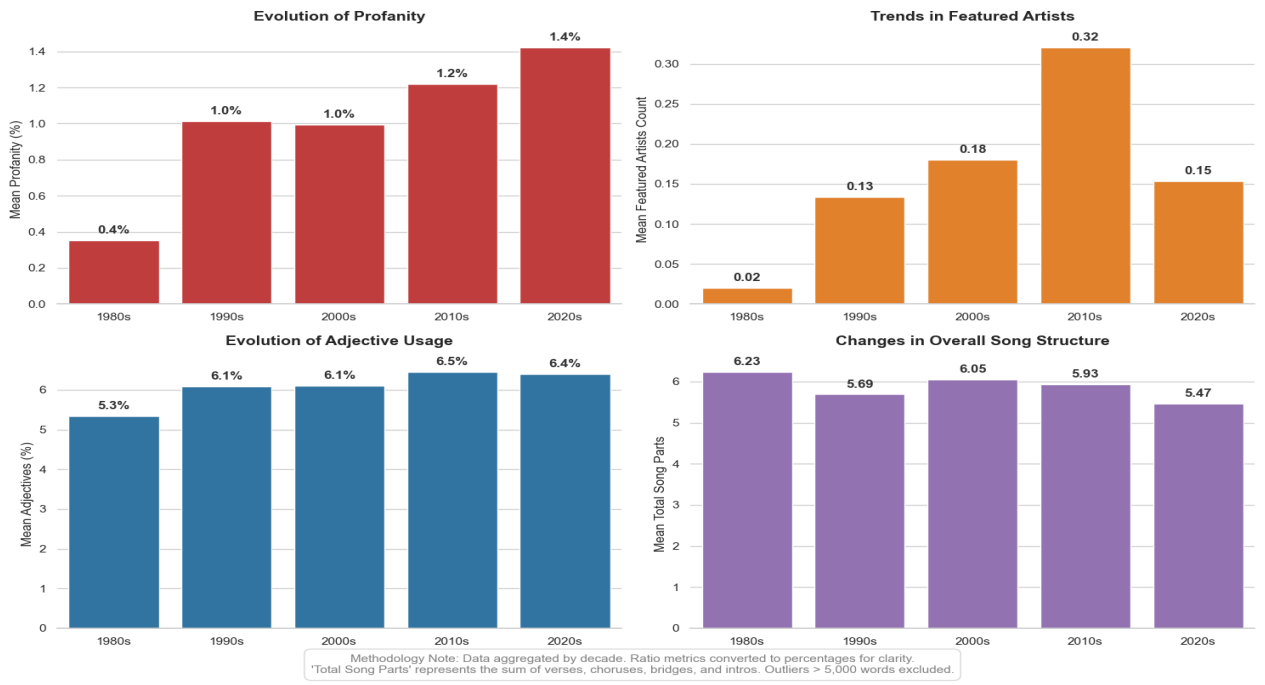


ניתן לדחות את השערת האפס רק עבור בלוז וקנטרי:
יש הבדל מובהק במדד העושר הלשוני בז'אנרים אלה בין תקופות העשורים הנ״ל.

שאלה 3 – משתנים נוספים:

מלבד השינויים באורך השיר ובמורכבות השפה, כיצד התפתחו המאפיינים הסגנוניים (כגון שימוש בשפה בוטה מול כתיבה תיאורית) ורמת שיתופי הפעולה (Featured Artists) במוזיקה הפופולרית על פני חמישה עשורים?

Figure 7: **Longitudinal Analysis of Musical Features by Decade (1980s-2020s)**



ניתן לזהות עלייה בבוטות המילים ומילות תואר לאורך השנים במקביל למגמת ירידה באורך מבנה השיר (חלוקה לחלקים שונים) (גרף 7)

סיכום לשאלת מחקר עד כה

1. מהנתונים הללו אין תצפיות המרמזות על ירידה באיכות המילים של השירים עם השנים, ואם כבר להפך מתקבלת תמונה הפוכה כאשר מתייחסים למדד העושר הלשוני לפי חלק מהז׳אנרים (בלוז וקנטרי) אנו צופים בעלייה מובהקת בעושר הלשוני.
2. צפינו בתופעות נוספות המאפיינות עשורים שונים במוזיקה הפופולרית בניהם:

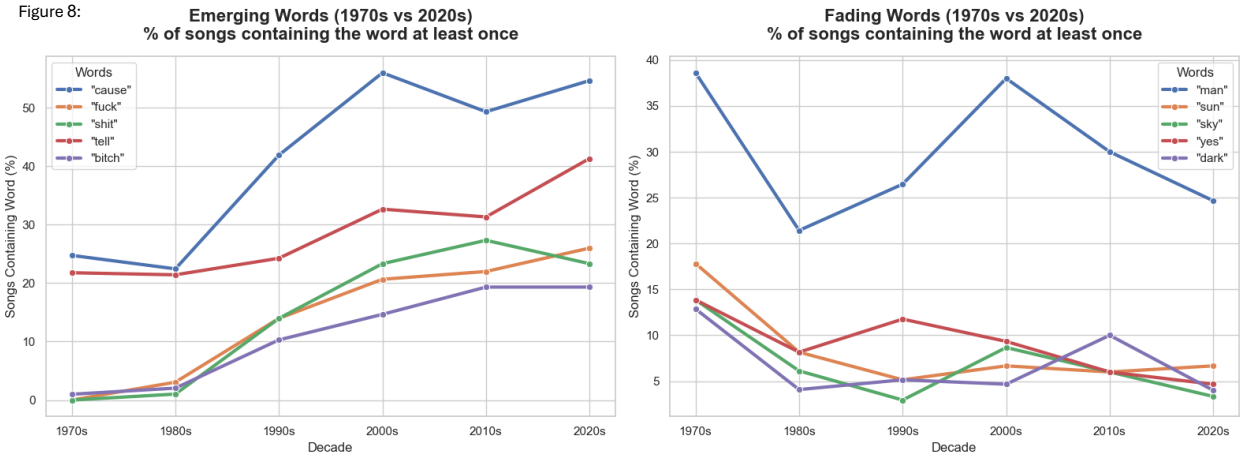
- אין שינוי מהותי באורך השירים.
- ישנה עלייה מתמדת בשימוש במילות גנאי – ניתן לייחס זאת לעלייתו של ז׳אנר ה׳היפ-הופ׳, וכניסתו והשפעתו על מוזיקת המיינסטרים.
- שינוי מהותי במספר אומני האורח הממוצעים לשיר - ניתן לפרש תצפית זו כעלייה בהתמסחרות האומנים השונים וניסיון להגיע לקהלים והאזנות רבות יותר.
- ירידה מסוימת במספר חלקי השיר, לשיר ממוצע.
- ישנה עלייה מסוימת בשימוש במילות תואר.

שאלת העמקה - אילו מילים "נכחדו" / "נוצרו" במהלך העשורים הנבדקים:

נבצע ניתוח של הנתונים לפי שתי שיטות:

1. אחוז השירים בהם מופיעה המילה לפחות פעם אחת.

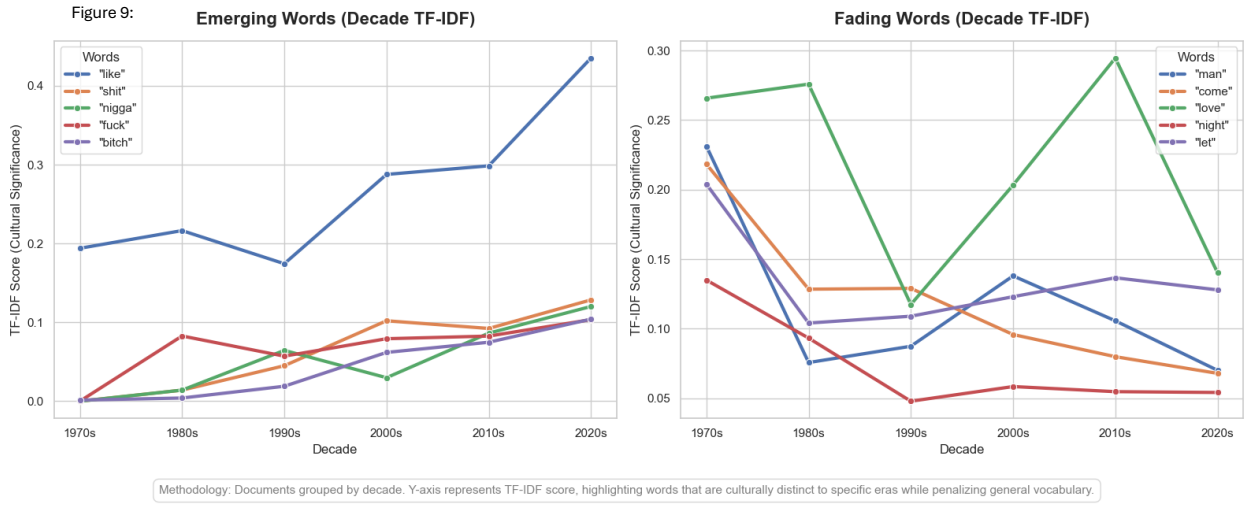
Figure 8:



2. ייחודיות כלל המילים בשירים בכל עשור בנפרד ביחס לעשורים אחרים, בעזרת מודל TF-IDF.

$$\text{TF-IDF Score} = \left(\frac{\text{Count of word in the decade}}{\text{Total words in the decade}}\right) \cdot \log\left(\frac{\text{Total number of decades}}{\text{Number of decades containing the word}}\right)$$

Figure 9:



מסקנות ביניים:

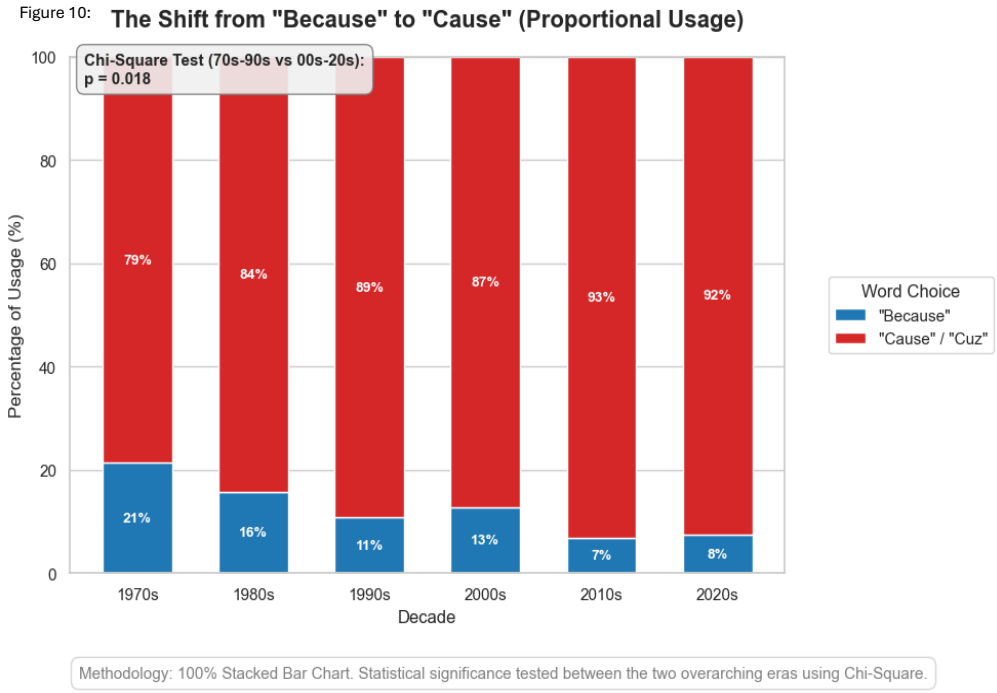
- ההבדל בין שני המבחנים מייצר לנו תוצאות שונות משום שבשיטה הראשונה אין משמעות לחזרה של המילה הנבדקת בתוך השיר (בינארי), לעומת השיטה השנייה שסוכמת את כלל המילים בכל עשור ולכן יש משמעות לכמות החזרות של מילה בתוך השיר.
- בגרפים של המילים ש"נוצרו" (Emerging Words)- אפשר לזהות מילים גסות שנמצאות במגמת עלייה עם העשורים, תצפית שמתאימה למבחן האבולוציה של הגסויות (figure 7).

ניתוח של המילה Cause:

העלנו השערה שעלייתה של המילה Cause (figure 8) נובעת משימוש מוגבר של הקיצור cause ונגזרתו (cuz) מתוך המילה Because ומהווה תחליף שלה בפועל.

נצפה לראות ירידה יחסית של המילה Because ביחס למילה Cause על פני העשורים.

Figure 10:



ניתן לראות שבהחלט יש ירידה יחסית של המילה Because והיא מובהקת סטטיסטית לפי מבחן חי בריבוע (p-value = 0.018)

- לפי נתונים אלה לא ניתן להסביר את ההקשר של המילים והאם הם מהוות תחליף מדויק לאותם הקשרים תחביריים.
על מנת להוכיח זאת יש לבצע מחקר המשך המנתח את הקשרי המשפטים ומשמעות השימוש במילה Cause כתחליף תחבירי ל Because.

העמקת בונוס – LLM's

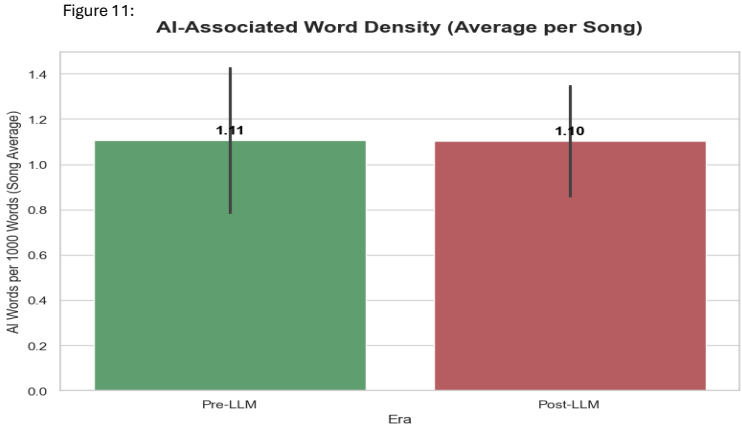
האם ניתן לראות עלייה יחסית במילים שנכתבות ע"י מודלי שפה של בינה מלאכותית בשירים שנכתבו בשנים האחרונות.

מתוך המחקר: **Delving Into PubMed Records: How AI-Influenced Vocabulary has Transformed Medical Writing since ChatGPT** **מאת** **Kentaro Matsui**
נראה שיש נטייה של מודלי ה-LLM השונים, להשתמש במילים מסוימות בתדירות גבוהה יותר בהשוואה לשימוש במילים אלו ע"י בני אדם.

לקחנו את רשימת המילים (125 מילים) שלפי המחקר מועדפים לשימוש ע"י מודלי הLLM ובדקנו את תדירותם ביחס לשתי קבוצות:

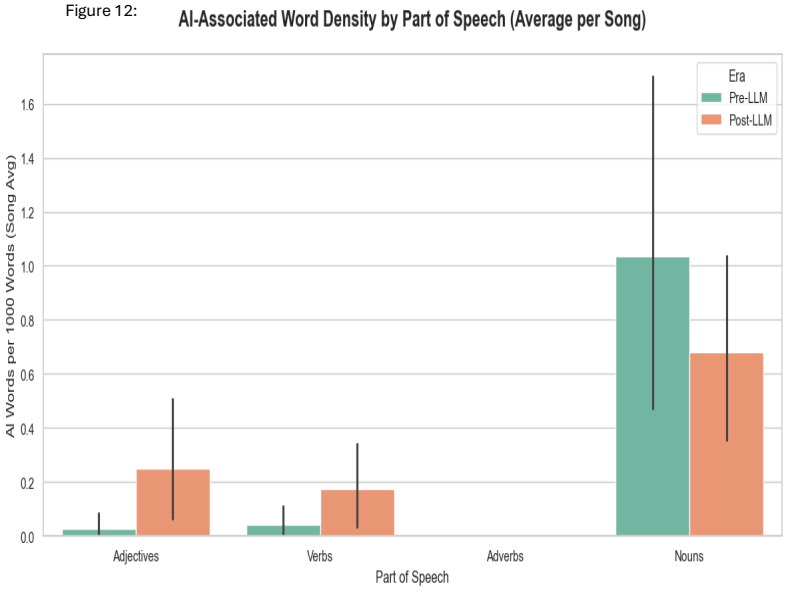
- Pre - LLM (2018-2022)
- Post - LLM (2023-2026)

Figure 11:



ניתן לראות שאין עלייה משמעותית במילים אלו בין הקבוצות ואם כבר ישנה ירידה עם השנים.

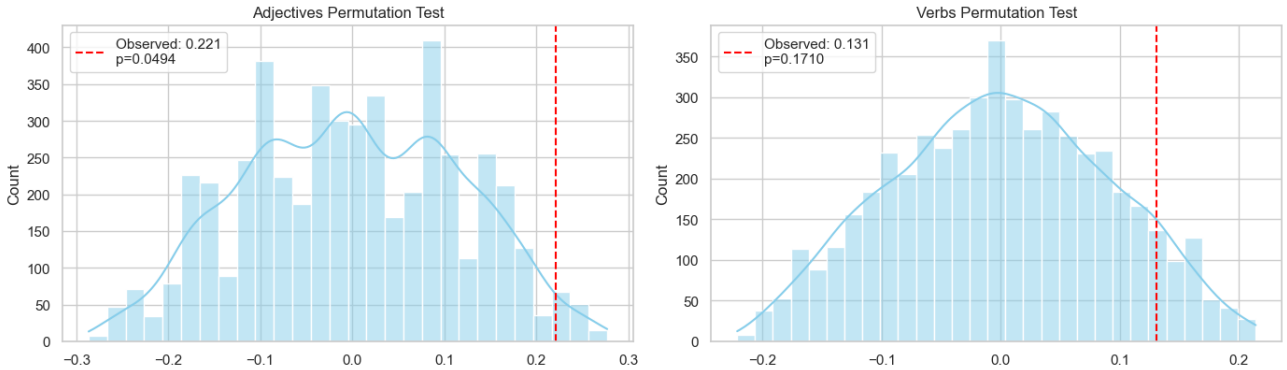
Figure 12:



אך לאחר ניתוח מעמיק המפרק את המילים ברשימה לתתי קבוצות בהם סוג המילה, ניתן להבחין במגמות מעניינות בניהם:

- אין בכלל מילות "תואר הפועל" (Adverbs) בכל הדאטה שלנו.
- ישנה עלייה בשימוש במילים מסוג "פועל" (verbs) ו "שמות תואר" (Adjectives).
- והנקודה ה "מפתיעה" היא שישנה ירידה ב "שמות העצם" (Nouns) והיא שגרמה לירידה שנצפתה בסך סוגי המילים (figure11)

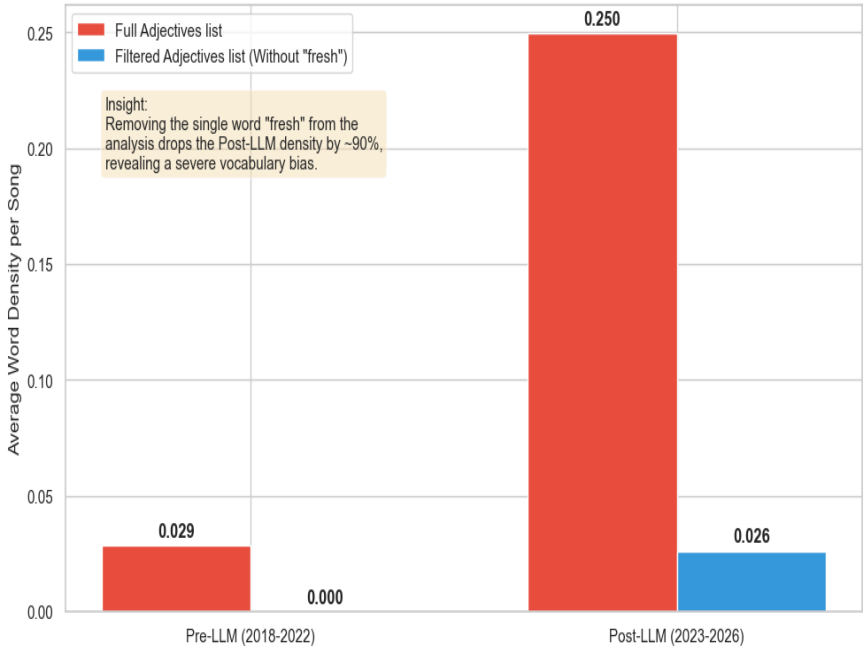
Figure 13: Permutation Test: Stability Analysis



לאחר מבחן פרמוטציות המערבב את השירים ביחס לשיוך לטווח השנים שלהם צפינו כי אין מובהקות לעלייה בשימוש במילות AI בפעלים (Verbs) אך כן ישנה מובהקות יחסית בתארים (Adjectives).

- שיטת החישוב שבצענו עבור בדיקות אלה היא בשיטת "צפיפות Ai ברמת השיר" – כלומר לכל שיר בנפרד מחשבים את אחוז השימוש שלו במילים מתוך רשימת הAi מתוך כל מילות השיר וממצעים את כל הריכוזים של כל השירים מאותה קבוצת שנים. בשיטה זו אנו מנטרלים אפקט של השפעה משמעותית של אורך השיר על הממוצע הכללי (המטרה לבחון האם השיר השתמש בLLM ולא לתת לגודלו להשפיע על כלל הקבוצה).
- לא ניתן להסיק מתוצאות אלה האם יש שימוש בLLM בשירים החדשים וגם לא ניתן לכמת כמה השתמשו אם בכלל. ניתן רק לומר שמילות תואר המזוהות עם הAi נמצאו בעלייה בתקופה הנ"ל.

Figure 14: Sensitivity Analysis: The Impact of Lexical Bias (The "Fresh" Effect)



אך לאחר בחינה מעמיקה של המילים שתרמו לעלייה בצפיפות התארים בעידן Post-LLM, התגלה כי המילה 'fresh' היא המקור המרכזי והכמעט בלעדי לזינוק בנתונים. להערכתנו, מדובר במצב שבו מילת סלנג, שצברה פופולריות טבעית בשירים בשנים אלו, הופיעה במקביל גם ברשימת המילים המזוהות עם- AI מה שיצר הטיה. על כן, בהתבסס על המדגם שניתחנו, המסקנה היא שלא קיימת עדות מהותית לכך שמודלי שפה (LLMs) השפיעו על אוצר המילים ועל סגנון הכתיבה של שירים פופולריים בתקופה הנבדקת.

שלב ניקוי הדאטה

כבר בשלב יצירת הדאטה נתקלנו במספר בעיות בניהם: חזרה של שירים דרך ז׳אנרים שונים, הופעת גרסאות ״קאבר״ לאותם שירים, הופיעו טקסטים שאינם שירים ״קלאסים״ כמו תמליל סרטים, שירים שאינם באנגלית. התגברנו על בעיות אלה דרך הגדרת ספים (threshold) לייבוא השירים כמו: שנת הוצאה, אורך שיר מינימאלי ומקסימאלי.

דיונים ומסקנות

- **מדגם** - מדגם המחקר אינו מייצג את האוכלוסייה. השירים מכל קבוצה שהגדרנו הינם השירים הפופולריים ביותר לפי פלטפורמת Genius ולא נבחרו בצורה רנדומלית, בנוסף ישנם רק 30 שירים לכל תת קבוצה ובסך הכל פחות מ800 שירים.
- **חלוקת הז׳אנרים** - ממצא מעניין העולה מהתפלגויות ה Null-במבחני הפרמוטציות (figure 6) הוא קיומה של התפלגות דו-דבשתית (Bimodal Distribution) בז'אנרים הבלוז והקאנטרי. תופעה סטטיסטית זו מצביעה בסבירות גבוהה על הימצאותן של תת-קבוצות מובחנות בתוך אותם ז'אנרים (לדוגמה, הבדלים אפשריים בין 'קאנטרי קלאסי' ל'קאנטרי-פופ'), אשר מתנהגות שונה מבחינה לקסיקלית ויוצרות רעש. ממצא זה ממחיש בעייתיות מסוימת בחלוקה לז׳אנרים השונים, זו מעין חלוקה שרירותית ללא מאפיינים אמיתיים בחלק מהמקרים.
- **LLM** - בשלב הזה השתמשנו ב125 מילים שנותחו כמילים בהם הLLM מבצע שימוש נרחב, אך המחקר ממנו לקחנו מילים אלה בעייתי לשימושינו משום שהוא מתייחס לניתוח מילים מתוך מאמרים מדעיים ולא מתוך קבצים ספרותיים או פואמות. ולכן מלכתחילה רשימת 125 המילים אינה מתאימה בהכרח לשירים, ולא בטוח שאלו המילים האופייניות לLLM בהקשר של כתיבת שירים.
- **שאלת המשך** - האם שירים פופולריים יותר הם פחות איכותיים (יותר טיפשים)?
ניתן לבדוק זאת על ידי שימוש בנתונים מתוך הדאטה-בייס של פלטפורמת השמעת שירים כמו ספוטיפיי/ יוטיוב. שבהם יש מדד צפיות/ האזנות ולבצע ניתוחים סטטיסטיים של משתנה זה בשילוב הדאטה מהפרויקט שלנו. לא נוכל להסתפק בנתונים הקיימים בפלטפורמת Genius. למרות שישנם משתנים של ליקיים לשירים וכמות תגובות לכל שיר, משתנים אלו אינם מעידים על פופולריות אמיתית של השירים, מפני שGenius אינה פלטפורמה המאפשרת האזנות לשירים אלא לקריאת הטקסט וקריאת פרשנויות על הטקסט בלבד.