עבודה סמינריונית במסגרת תואר שני

אוניברסיטת בר אילן

בנושא:

אפיון סגנוני לפירוש רש"י לתלמוד וזיהוי פרשנים אחרים

שם הקורס: מבוא למדעי הרוח הדיגיטליים

מספר קורס: 843595401

שם המרצה: דוקטור גילה פריבור

מגיש: ניר דעוס 315782987

תוכן עניינים

מבוא

פירוש רש״י לתלמוד מהווה את אחת היצירות החשובות ביותר בעולם הרוח היהודי. האופי הפשטני והתמציתי מהווים מקור חשוב לכל לומד מתחיל או מנוסה בהבנת דברי התלמוד. למרות שקיימים פירושים לתלמוד גם לפני זמנו של רש״י, כולם נדחו מפניו ומאז הדפסת התלמוד לראשונה לא הייתה מהדורה שלא כללה את פירושו. [1] במרכז המחקר על פירוש רש״י לתלמוד עומדת שאלת ההיקף – לאלו מסכתות כתב רש״י את פירושו?

ידוע לנו שרש״י פירש את רוב המסכתות בתלמוד ואולי אף את כולו. אך מקצת הפירושים הנדפסים כיום המיוחסים לו אינם שלו וחלקם נמצאים בזהות מסופקת. עבודה זו מתמקדת באפיון סגנון פירוש רש״י במסכתות בהן ברור שהוא הפרשן בהשוואה למול הפירושים שאינם שלו. כתוצאה מכך יהיה ניתן לקבל תובנות באשר לזהות הפירוש השנוי במחלוקת הייחוס.

[1] י. תא שמע, הספרות הפרשנית לתלמוד, א׳, ירושלים, תש״ס, עמ׳ 41.

מטרות המחקר

מטרת המחקר היא לנתח תכונות סגנוניות של פירוש רש״י לתלמוד במסכתות הידועות שהוא הפרשן, להשוות את המדדים הסגנוניים הללו לפירושים במסכתות שייחוס פירושן שנוי במחלוקת, ולהכריע בשאלת הייחוס.

שאלות המחקר:

1. האם ניתן באמצעות כלים של ניתוח סגנוני לזהות הבדל בין הסגנון של רש״י לסגנון של פרשנים אחרים?
2. האם ניתן באמצעות כלים של ניתוח סגנוני לסווג את הפירוש במסכתו תענית, הוריות ומעילה כקרוב לסגנונו של רש״י או לא?

סקירה ספרותית

סקירה זו נועדה לתת רקע רלוונטי לפירוש רש״י לגמרא – חשיבותו, אפיון סגנוני ראשוני, מחלוקות על מסורת הייחוס, ויחס הספרות הרבנית והמחקרית אליו. בנוסף לכך, סקירה זו תעסוק בהתפתחות השיטות לניתוח סגנוני על מנת לייחס טקסט אל מחברו ובפרט במחקר כתב יד דתיים היסטוריים.

סגנונו של רש״י בפירושו לתלמוד

פרופ׳ יונה פרנקל בספרו ״דרכו של רש״י בפירושו לתלמוד הבבלי״ מאפיין את סגנונו של פירוש רש״י לגמרא. לדבריו, רש״י רואה את עצמו כמורה המדריך את הלומד צעד-צעד בלימוד הגמרא. אך פירושו אינו לתלמיד מתחיל, אלא לתלמיד מנוסה שכבר למד גמרא ויודע מונחים בעברית וארמית. לכן, בפירושו, רש״י אינו מנסח שוב את דברי הגמרא והוא מניח שיש לתלמיד זיכרון של הסוגייה גם בהקשר הכללי שלה בכל התלמוד.[2]

רש״י נצמד לטקסט ובדרך כלל אינו מתייחס לסוגיות מקבילות או הקשרים סותרים בגמרא, כפי שנוהגים לעשות בעלי התוספות בסגנונם המעמיק יותר. רש״י בא לבאר את הגמרא כולה ולא כקטעים נבחרים, כפי שמצוי בפירושם של אחרים. סגנונו מתאפיין בקצרנות תמציתית ביותר ולעיתים פירושו כולל מילים ספורות בלבד[3] . וכך מאפיין אותו רבי יצחק די-לאטיש, רב צרפתי שחי במאה ה-14[4]:

> ״וראש כל החיבורים שנתחברו דרך פירוש הם פירושי הרב רבינו שלמה בר יצחק הנזכר. ואם רבו הלוחמים עליו כלי זיינו עליו. **ותשובתו בתוך דבריו,** כלם נכוחים למבין. אין מעלתו נכרת רק ליחידים. כי **במלה אחת יכלול פעמים תירוצין של חבלי קושיות**.״ [הדגשות שלי – נ״ד]

חכמי ספרד האחרונים התייחסו לחשיבות הנסתרת שבכל מילה שבחר רש״י. וכך כתב ר׳ יצחק קנפנטון, מחכמי ספרד במאה ה-15[5]:

> ״מנהגי רש״י ושיטתו היא שלא לדבר דבר ושלא להוציא מלה בלשונו שלא לצורך״

האר״י קבע שיש לדקדק אפילו באותיותיו של רש״י[6]:

> ״ודקדק מאוד בלשונו שרמז כמה חידושים בשינוי אות״

לסיכום, פירושו של רש״י הוא מדוקדק, מתומצת ונצמד לטקסט המקביל בגמרא. לעיתים נדירות נראה שרש״י מאריך או מזכיר סוגיות שאינם על דף הגמרא המקביל לפירוש. סגנונו של רש״י הוא בעל מאפיינים ייחודים וסגנונו עקבי לאורך פירושיו.

---

[2] י׳ פרנקל, דרכו של רש״י בפירושו לתלמוד הבבלי, ירושלים, תשל״ה.

[3] מ׳ גרוסמן, חכמי צרפת הראשונים, ירושלים, תשנ״ז, עמ׳ 217.

[4] די-לאטיש יצחק, הבלין, סדר הקבלה, עמ׳ 147.

[5] י׳ קנפנטון, דרכי התלמוד, מהדורת י״ש לנגה, ירושלים תשמ״א, עמ׳ 59.

[6] חיים יוסף דוד אזולאי, שם הגדולים, מערכת גדולים, ערך רש״י, עמ׳ 203.

מחלוקות בייחוס הפירוש בתלמוד

הפירוש למהדורות הנדפסות כיום ברובו מיוחס במסורת לרש"י. אך יחד עם זאת, מקצתם אינם שלו. לא נשתמרו פירושיו של רש"י, אם היו כאלה, למסכתות נדרים, נזיר, מועד קטן, בבא בתרא (מדף כט ע"ב) ומסכת מכות (מדף כ ע"א). כנראה שמפרשים אחרים השתרבבו בטעות על ידי מעתיקים כפירוש רש"י במקומו. פירושו למסכת מועד קטן נתגלה רק מאוחר יותר בשנת תשכ"א[7], הפירוש המצוי היום בגמרא מעורב מדברי רש"י ותוספות מאוחרות יותר[8].

בנוסף לכך ישנם ספקות באשר לייחוס הפירוש למסכתות תענית, הוריות ומעילה:

1. **הפירוש למסכת תענית**

הרב צבי הירש חיות, שחי במאה ה-19, חלק על שיוכו של הפירוש על תענית לרש"י במאמרו "אמרי בינה" סימן ט':

" ראיתי להרבה חכמי ישראל אשר נסתפקו, אם פירוש"י אשר נמצא אצלנו על מסכת תענית, בא לנו באמת מן אבי התעודה הפרשן המפורסם רש"י זצ"ל, או פירוש זה רק מיוחס אליו לבד... מצאתי ראיות נכונות וברורות אשר יתנו עדיהן ויצדקו, כי **אין אפשרות כלל שיהיה פירוש זה מן רבינו הגדול רש"י** ... "
[הדגשות ודילוגים שלי – נ"ד]

מבין הסיבות שהביא:

א. סתירות מהותיות בין הפירוש בתענית לרש"י במקומות אחרים.
ב. נוסח משונה של בעלי התוספות על תענית שמעיד על כך שלא היה זמין להם פירוש רש"י.
ג. נוסח הפירוש בתענית מאריך יתר על המידה ביחס לרש"י (מעתיק מקרא שלם, אינו מדוקדק ועוד).

לעומת זאת, החיד"א בספרו שם הגדולים בערך "רבי שלמה יצחקי" מתייחס למחלוקת וסובר שזהו דווקא כן פירושו של רש"י:

"הרב הנזכר במשנה לחם הוסיף דפירוש מס' תענית אינו מרש"י... האמת אחר זמן ראיתי דמפירוש דברי רבינו ישעיה והריטב"א ורבינו בצלאל בחדושיהם למסכת תענית **נראה בברור דפירוש רש"י שלנו במסכת תענית הוא פירש"י** אלא שמרוב העתקות נפלו כמה ט"ס. וכן כתבתי בעניותי.." [ההדגשות שלי – נ"ד]

החיד"א מסיק מפירושי הראשונים הנ"ל שהפירוש לתענית הוא אכן של רש"י ופותר את כל הסתירות בטענה של טעות סופר.

---

[7] א' קופפר, פירוש רש"י למסכת מועד קטן, ירושלים תשכ"א.

[8] יואל פלורסהיים, פירושו של רש"י למסכת מועד קטן, חוברת תרביץ, גיליון ניסן-סיוון תשמ"ב.

2. הפירוש למסכת הוריות

פרופ׳ יעקב נחום אפשטיין במאמרו ״המיוחס לרש״י בהוריות״[9], סובר שפירוש רש״י להוריות שייך לרבינו גרשום (רגמ״ה). פרופ׳ אפשטיין מתייחס, בנוסף לסתירות בין פירושים אחרים של רש״י, לאורך הפירוש, למילים שרש״י לא רגיל להשתמש בהן ולחוסר התעוזה לפסוק פסק הילכתי שאינו אופייני לרש״י.

הוכחות נגדיות לטענותיו של פרופ׳ אפשטיין הן כתבי הרא״ש ורמ״ה, הם רבי אשר בן יחיאל ורבי מאיר בן טורדוס, שמתייחסים באופן ברור לפירוש זה כפירושו של רש״י. אך לטענתו של הפרופ׳ אין לסמוך עליהם:

> ״רמ״ה מביא כמה דברים מפירוש שלנו זה בשם ״רש״י״ או ״יש אומרים״.
> והרא״ש בתוספותיו להוריות אף הוא מזכיר כמה דברים מפירושנו בשם ״רש״י״. אבל עדותו זו של רמ״ה הספרדי (מטולידא), שלא ידע יפה כל יחסם של פירושי רש״י לתלמוד, **אינה מוסמכת**. והרא״ש נמשך כאן אחרי הרמ״ה וייחסו אף הוא לרש״י. וראיה לדבר הוא, שברוב המקומות שהוא מזכיר בהוריות דברי רש״י — לקוחים דבריו מדברי הרמ״ה״ [הדגשות שלי – נ״ד]

כלומר, אין לסמוך על עדותו של הרמ״ה, כיוון שלא ידע באופן ברור את הייחוס של הפירושים לתלמוד. וכיוון שהרא״ש הסתמך עליו, גם על עדותו אין לסמוך.

לעומת שיטתו של פרופ׳ אפשטיין, המסורת לא חולקת על היחוס של הפירוש למסכת הוריות לרש״י, ולא נמצא אף מקור רבני שתומך בכך. יתר על כן, הרב בצלאל דבליצקי דוחה את טענותיו[10]:

> ״בעת האחרונה יצאו לערער על פירוש רש״י למסכתנו, לבטל חזקת ראשונים ולקבוע שפרוש זה לא מרש״י יצא... מצער עוד יותר כי נעלם מעין העוסקים בדבר כי רבותינו הראשונים מביאים מן הפירוש שנשמר לנו בשמו של רש״י...
> מן המעטים אשר עסקו בשאלה זו היחיד שנתן נימוקים לדבריו הוא י״נ אפשטיין״.
> [הדילוגים שלי – נ״ד]

בהמשך דבריו מתייחס הרב בצלאל לכלל הטענות בדבר חוסר האחידות בנוסח פירוש רש״י, ובכללם טענות פרופ׳ אפשטיין וגם לטענות שהובאו נגד הפירוש למסכת תענית:

> ״אפשר שהפתרון לחוסר האחידות הסגנונית המתגלה לעיתים בפירושי רש״י, מצוי במהדורות שעשה רש״י לפירושיו.״

---

[9] י.נ. אפשטיין, המיוחס לרש״י בהוריות, 1995.

[10] בצלאל דבילצקי, פירוש רש״י להוריות עפ״י כת״י, 2005.

הרב בצלאל פותר את כל טענות חוסר התאימות הסגנוני למהדורות שונות של הפירוש, שלא נשתמרו כולן, בכך שניתן לטעון שהפירוש להוריות והפירוש לתענית הם ממהדורה מוקדמת יותר[11] ולכן סגנון הפירוש שונה.

3. הפירוש למסכת מעילה

הרב מלאכי הכהן, שחי במאה ה-18, כתב בספרו "יד מלאכי" בשם הרב אהרון קאיידנובר בספר "ברכת הזבח", שפירוש זה אינו מרש"י[12]:

> "פירוש מסכת מעילה, **אינו משל רש"י**, אלא מאחד מאיזה תלמידיו, דוק ותשכח כנ"ל. [=דייק ותמצא כנאמר לעיל. נ"ד]"

וכנראה שהייתה מסורת בידם כי לא הביאו הוכחה לדבריהם. אולם, הרב חיים חזקיהו, שחי במאה ה-20, הביא בספרו "שדי חמד" ראיה מדברי הרב יוסף קארו ("מרן") שמתייחס לפירוש זה כשל רש"י[13]:

> "ותמיה לי... דפירוש שמכנה אותו מרן כבוד מעלתו, בשם רש"י הוא **בוודאי מרש"י עצמו.** שאם לא כן, היה מכנה אותו בלשון 'המפרש'". [דילוגים והדגשות שלי – נ"ד]

לסיכום, ניתן לראות שהייחוס לרש"י אינו חד משמעי במסכתות אלה. לכן במסגרת העבודה הזו אנסה להביא סעד לאחד הצדדים ולנסות לסווג את הפירושים השנויים במחלוקת לרש"י, או דווקא לשלול את הייחוס. טענות הנוגעות לסתירה מהותית לא יבדקו במסגרת מחקר זה כיוון שהעבודה מתמקדת באפיון ניתוח סגנוני של המחבר. לגבי הטענה למהדורות שונות, אפשר שישפיע על אורך הפירוש אבל הסגנון הכללי של הפרשן (לדוגמא, שכיחות מילים וביטויים) לא משתנה.

שיטות מחקר קודמות לזיהוי מחברים במחקר המודרני

בעיית יחוס טקסט למחבר כשאין הוכחות למקורו התפתחה עם השנים לתחום מחקר הנקרא "Stylometry" – ניתוח סגנוני. בשנים האחרונות תחום זה קיבל תנופה משמעותית עם כניסתם של כלים חישוביים לניתוח סגנוני ממוחשב כמו דיקטה, Voyant וכלי למידת מכונה וסטטיסטיקה. שילוב של בינה מלאכותית ומודלים מתקדמים כגון רשתות נוירונים מביא איתו יכולת חזקה לנתח טקסטים רבים וארוכים ולהגיע למסקנות במהירות.

באמצעות ניתוח סגנוני אנחנו מבקשים למצוא את המאפיינים הסגנוניים הייחודיים, שקשה להעתיק או לזייף, של כותב מסוים. המדדים הבסיסיים שמקובל להשתמש בהם כוללים: שימוש במילות קישור, אורכי משפטים ומילים, שימוש במילים לא שכיחות, שימוש בביטויים ועוד. באמצעות שיטות אלה ניתן

---

[11] לנושא האם היו מהדורות שונות לפירוש רש"י: אפטוביצר, לתולדות פירוש רש"י לתלמוד, שנת תש"א, עמוד שכ. חיים יוסף דוד אזולאי, שם הגדולים, מערכת גדולים, שנת התקנ"ח, ערך רש"י, עמ' 203. ואכמ"ל.

[12] מלאכי הכהן, יד מלאכי, שנת התקכ"ז, עמוד סו.

[13] חיים חזקיהו, שדי חמד, כללי הפוסקים, סימן ו'.

להבחין בסגנון ייחודי אפילו שבאופן כללי הכתיבה נראית לנו זהה. ניתן גם לקרוא לשיטות האלה "קריאה מרחוק", כיוון שבניגוד לקריאה הרגילה, "קריאה מקרוב", שמים לב לפרטים הסגנוניים הקטנים האלו. ניתן להשתמש בכלים של ניתוח סגנוני כדי להוכיח זיופים של טקסטים או לקביעת המחבר של טקסט היסטורי.

המחקר הקלאסי הראשון בתחום זה הוא של פרדריק מוסתלר ודיוויד וולס[14] בשנת 1964 שניתחו מכתבים אנונימיים מהמאה ה-18 בניסיון לגלות את זהות המחבר מתוך רשימה פוטנציאלית. פריצת הדרך באותה העת הייתה השימוש בתדירות מילות הקישור, שלא מוסיפות לתוכן הטקסט, בתור מדד סטטיסטי לקביעת ההסתברות למחבר הנכון וזיהוי הסגנון הייחודי של המחבר על סמך אלה בלבד. מחקר זה הוכיח שאין צורך להבין את תוכן הטקסט, אלא מספיק למדוד את המאפיינים שלו.

במחקר מהשנים האחרונות, במאמר מאת פרופ' משה קופל וירון וינטר[15], הודגם כיצד ניתן לקבוע האם שני טקסטים אנונימיים בסגנונות ואורכים משתנים מהאינטרנט נכתבו על ידי אותו מחבר. במאמר הנ"ל הוכח שבאמצעות שימוש בשיטות סגנוניות (תדירות מילים, אורכי משפטים, שימוש במילות קישור) ניתן לקבל "טביעת אצבע" סגנונית של הכותב, ללא תלות באורך, שפה או בנושא הטקסט.

במחקר היהודי, קיים עניין ביישום כלי ניתוח סגנוני על כתבים מסורתיים-דתיים. פרופ' משה קופל חקר ופיתח שיטות לזיהוי מחברים גם בטקסטים מסורתיים ועסק גם בהבחנה בין סגנונם של פוסקים ורבנים מתקופות שונות.

לדוגמא, במאמר שפרסם בקובץ התורני "ישורון" (אלול תש"ע), "זיהוי מחברים בשיטות ממוחשבות: המקרה של גניזת חרסון"[16] הוא עסק בזיהוי מכתבים המיוחסים לרב שניאור זלמן מלאדי (רש"ז) שנמצאו בגניזה בעיר חרסון באוקראינה. באמצעות ניתוח סטטיסטי של שכיחות המילים בכתבי הרש"ז המקוריים ניתן ליצור מודל מסווג שמייצג את סגנון הכתיבה שלו. לאחר מכן ניתן להשוות את מודל השכיחות למול שכיחות המילים ממכתבי הגניזה בחרסון ולהגיע למסקנה על פי רוב הדמיון בשכיחויות.

לדוגמא, טבלת חמשת המילים הנפוצות ביותר בכתבי רש"ז:

מאפיין	שכיחות רש"ז-אוטונטי	שכיחות חרסון-אחרים	שכיחות חרסון-רש"ז	תוצאה: שכיחות חרסון-רש"ז דומה ל-
זאת	0.26%	0.26%	0.00%	שכיחות חרסון-אחרים
זו	0.23%	0.00%	0.02%	שכיחות חרסון-אחרים
ולא	0.52%	0.08%	0.13%	שכיחות חרסון-אחרים
לא	0.76%	0.36%	0.29%	שכיחות חרסון-אחרים
שלא	0.52%	0.06%	0.07%	שכיחות חרסון-אחרים

טבלה 0- טבלת שכיחויות מילים מתוך מחקר של פרופ' קופל.

[14] Mosteller, F., & Wallace, D. L. (1964). *Inference and disputed authorship: The Federalist*. Addison-Wesley.

[15] קופל, משה ווינטר, ירון. " Determining if Two Documents Are Written by the Same Author", 2014.

[16] קופל, משה. "זיהוי מחברים בשיטות ממוחשבות: המקרה של גניזת חרסון", ישורון, אלול תש"ע.

בטבלה זו ניתן לראות בעמודה הראשונה את המילים הנפוצות ביותר בכתבי הרש״ז האותנטיים. בעמודה השנייה ניתן לראות את השכיחות של המילים הנפוצות ביותר בכתבי הרש״ז האותנטיים. בעמודה השלישית, את השכיחות של המילים בכתבים שאינם של רש״ז מתוך גניזת חרסון. ובעמודה הרביעית, את השכיחות של מילים מתוך הכתבים המיוחסים לרש״ז בגניזת חרסון.

לאחר מכן, המרחק חושב בין עמודה 4 לעמודות 2,3 והתוצאה נכתבה בעמודה האחרונה. אם רוב המילים היו בעלות שכיחות קרובה לרש״ז האותנטי היינו יכולים להסיק שאכן המכבתים מגניזת חרסון מיוחסים לרש״ז.

אך לאור הממצאים האלו, הצליח פרופ׳ משה קופל להסיק שהמכתבים דווקא מזויפים ואינם מיוחסים לרש״ז, כיוון שרוב השכיחויות היו דומות לשכיחויות של ״חרסון-אחרים״.

סיכום

פירוש רש״י לתלמוד הוא בעל מאפיינים ברורים וסגנון עקבי. במאמר הזה אנסה להשתמש בכלי הניתוח הסגנוני שהוזכרו על מנת לקבל אפיון סגנוני של רש״י ובאמצעותו להבדיל או להשוות בין פירוש רש״י המוסכם לבין הפירוש בעל היחוס השנוי במחלוקת על מסכתות תענית, הוריות ומעילה. ובכך להוסיף ולתת סעד לצד במחלוקת של גדולי ישראל והחוקרים מהעידן המודרני על שיוך הפירוש למסכתות תענית, הוריות ומעילה.

מתודולוגיה

המידע

הפירוש לגמרא נלקח מאתר ״ספריא״. ״ספריא״ הוא ארגון ללא כוונת רווח המוקדש לעיצוב עתיד הלימוד היהודי באופן פתוח ושיתופי בין משתמשים. הוא משמש כספרייה דיגיטלית נגישה לציבור הרחב, המכילה טקסטים יהודיים בעברית ובתרגומים שונים. הטקסטים הדיגיטליים מסייעים בהנגשת הספרות מעולם היהדות לקהל רחב יותר. לצורך העובדה, הורדתי את הטקסט בפורמט ״מהדורת וילנא״, שהיא המהדורה הרווחת בקרב לומדי הגמרא.

איור 1- הורדת פירוש רש״י מאתר ״ספריא״.

המידע מגיע כטקסט פשוט שמורכב מציטוט הגמרא והפירוש עליו. הטקסט מכיל גם סימני פיסוק ותווים. בתוך הטקסט, נמצאת כותרת באנגלית שמציינת את הדף, העמוד והשורה במהדורה המודפסת. בנוסף לכך, קיימת חלוקה באמצעות שורה חדשה לכל פירוש. החלוקה הזאת מסייעת לעיבוד סטטיסטי של המידע כאשר באמצעותה אפשר לחלק את הפירוש לחלקים קטנים יותר ולראות ממוצעים בין פירושים, דפים, עמודים, שורות ומסכתות.


Rashi on Chagigah Vilna Edition.txt

Open with TextEdit

```
Rashi on Chagigah
רש"י על חגיגה
Vilna Edition
https://he.wikisource.org/wiki/%D7%AA%D7%9C%D7%9E%D7%95%D7%93_%D7%91%D7%91%D7%9C%D7%99

רש"י על חגיגה



Daf 1a


Line 1


Daf 1b


Daf 2a


Line 1

מתני' הכל חייבין בראייה - במצות ראיית כל זכורך (שמות כג) שצריכים להתראות בעזרה ברגל:
חוץ מחרש שוטה וקטן - דלאו בני דעה נינהו ופטורין ממצות:
החיגר והסומא - כולה יליף מקראי בגמרא:
ומי שאינו יכול לעלות ברגליו - מירושלים לעזרה ובגמרא מפרש להו:

Line 2

אי זהו קטן כו' - אבל מכאן ואילך אף על פי שאינו חייב מן התורה הטילו חכמים על אביו ועל אמו לחנכו במצות:
שלש רגלים - הראוי לעלות ברגליו חייב הכתוב וכיון דגדול פטור מן התורה קטן לאו בר חינוך הוא:
```

איור2 - פירוש רש"י למסכת חגיגה שהורד מאתר "ספריא".

חילקתי את המידע הנדרש לקטגוריות:

1. מסכתות שרש"י הפרשן באופן מובהק – רשימה חלקית מכמה סדרים כדי להפחית הטיה:
   a. זרעים: ברכות.
   b. מועד: מגילה, פסחים, ביצה, סוכה, שבת, חגיגה, יומא.
   c. נזיקין: סנהדרין, בבא קמא, בבא מציעא.
   d. נשים: קידושין, גיטין.
   e. קדשים: זבחים.

   יש לציין שאף שרוב המסכתות הן מסדר מועד, רובן קצרות. בשביל לאזן ולמנוע הטיה הובאו מסכתות ארוכות מסדר נזיקין שגם מאזנות את סגנון השפה השונה בין הסדרים הנובע מהשוני בנושא הלכתי לעומת נושא משפטי.
2. מסכתות שרש"י אינו הפרשן באופן מובהק –
   a. נשים: נדרים, נזיר.
   b. מועד: פסחים (פירוש רשב"ם).
3. מסכתות בהן הייחוס אינו חד-משמעי –
   a. מועד: תענית.
   b. נזיקין: הוריות.
   c. קדשים: מעילה.

4. שאר המסכתות נפסלו בגלל שהפירוש הוא מעורב או קטוע. הטקסט לא נלקח בחשבון היות ואין דרך וודאית לברור את רש״י משאר המחברים ולכן שימוש בו עלול לעוות את התוצאות. לדוגמא מסכת מכות, מסכת בבא בתרא, מסכת מנחות, מסכת תמיד[17] ועוד.

יש לציין שעל אף המאמץ לגוון את המדגם, עדיין יכולה להיות הטיה מכיוון שהמסכתות אינם מחולקות בשווה בין הסדרים או באורכן. בנוסף, אין הרבה מסכתות שהפרשן עליהם הוא לא רש״י. פרשנים כמו הרא״ש, הרי״ף, רבינו גרשום או הריטב״א כתבו פירוש מעמיק וארוך יותר שמופיע בד״כ בסוף המסכת ולא לצד הגמרא ולכן אינם מתאימים לניתוח במסגרת מחקר זה. לכן, עלולה להיווצר הטיה כיוון שאוצר המילים מוגבל לנושא המסכתות היחידות שמפרשם לצד הגמרא אינו רש״י.

ניקוי הטקסט

לניקוי הטקסט השתמשתי בשפת פייתון. משימות הניקיון כללו:

1. הסרת סימני פיסוק ותווים.
2. הסרת הפניות בסוגריים, ואנגלית שמופיעה בתבנית של הטקסט (כותרות שורה, דף).
3. הסרת ״הדיבור המתחיל״ כיוון שאינו חלק מהפירוש.
4. הפרדה בין פירושים לפי שורות חדשות.
5. הפרדה בין דפים באמצעות הכותרות שמופיעות בטקסט מאתר ״ספריא״.

המצב הסופי של הטקסט, לפני עיבודו, הוא טבלה בה בכל שורה, שם הדף ופירוש אחד בלבד ללא סימני פיסוק.

המדדים

לאחר מכן בחרתי אילו מאפיינים לחלץ מהטקסט על מנת לאפיין את סגנון הכותב:

1. **מדדי אורך הפירוש**[18]:

מדדים אלו חשובים כיוון שרש״י ידוע כפרשן תמציתי, לכן נצפה לראות שפירושו קצר יותר בממוצע מפירושם של אחרים.

   a. **ממוצע אורך הפירוש במילים** – כדי לקבוע כמה ארוך הפירוש במילים.
   b. **ממוצע אורך הפירוש בתווים** –כדי להראות שימוש במילים מורכבות.
   c. **ממוצע אורך מילה בתווים** – כדי לראות הרגל שימוש במילים ארוכות.

2. **מדדי איכות הפירוש**:

מדדים אלו יעידו על כשרון הפרשן בשימושו במילות קישור רבות וברוחב אוצר המילים שלו.

   a. **ממוצע אחוז השימוש מילים ייחודיות** – כמה מתוך המילים בפירוש הם מילים ייחודיות בטקסט. מדד זה מעיד על הסגנוניות של הכותב ורוחב אוצר המילים.

---

[17] ר׳ בצלאל אשכנזי ,שיטה מקובצת, דף ל״ג, עמוד ב׳.

[18] ״פירוש״ הכוונה להערה פרשנית אחת של הפרשן ולא לכל הפירוש על המסכת או הגמרא.

b. **ממוצע מורכבות הפירוש** – מדד זה חושב באמצעות חישוב אחוז מילות הקישור מתוך כלל המילים בפירוש. כאשר נקודת ההנחה היא שפרשן מיומן ישתמש בפחות מילות קישור. המחקר נסמך על רשימה של מילות קישור בעברית וארמית (הובאו בנספח 1). מדד גבוה מעיד על מורכבות גבוהה וקושי לפרש באופן פשטני.

3. **שימוש במילים בארמית לעומת המקבילות בעברית** – מעידה על סגנון הכותב בהעדפתו לשפה. בהיעדר מילון ערבי-ארמי זמין ברשת (לביצוע שאילתות בקוד), המחקר מתבסס על רשימה מצומצמת של מילים שהם נפוצות בגמרא בעברית ובארמית מתוך ידע כללי וגם מהאינטרנט (הובאו בנספח 2). מדד גבוה מעיד על העדפת המילים הארמיות.
4. **ממוצע שכיחות שימוש בלע"ז לפירוש** – מילים מהשפה הצרפתית העתיקה שרש"י נוהג להשתמש בה. איתרתי בטקסט בקלות כי רש"י נוהג לכתוב "בלע"ז" לאחר מילה כזו ויופיע גרש באמצע המילה.
5. **טבלת שכיחות מילים** – כמו במקרה של "גניזת חרסון", לבחון את שכיחות המילים הנפוצות של רש"י למול מפרשים אחרים.
6. **ביטויים נפוצים ותובנות מ-Voyant** – מעידים על סגנון בשימוש בביטויים ששגורים בלשונו ועוד.

בשביל להימנע מהטיות שנובעות מאורך המסכת או אורך הדף הגדרתי את כל המאפיינים הסטטיסטים להיות ממוצע לפירוש אחד. כך נמנעתי מתוצאות מעוותות כאשר המסכת ארוכה או קצרה באופן חריג וכן באשר לאורך הדף. בשביל להימנע מהטיה שקשורה לתוכן הנלמד המסכת, השתמשתי במגוון מייצג מכל סדר (מועד, נזיקין) ומכל אורך של מסכת (ארוכה, קצרה) לכל קטגוריה של פרשנות (רש"י, לא-רש"י) בהתאמה למסכתות שאני רוצה לבדוק את ייחוסן – תענית – סדר מועד, הוריות – סדר נזיקין ומעילה – סדר קודשים.

בשימוש במילים וביטויים נושא המסכת עלול להיות בעייתי. לדוגמא, במסכת נדרים סביר להניח שהשורש נ.ד.ר יהיה שכיח. בשביל להימנע מהטיה זו אתעלם ממילים או ביטויים שקשורים לתוכן המסכת.

הכלים לעבודה

ניקוי הטקסט, הגרפים וכן חישוב רוב המאפיינים (1-8) הסטטיסטיים בוצע באמצעות סקריפט בשפת פייתון. בחלק מהמקרים עיבדתי את הפלטים עם אקסל. שפת פייתון נבחרה בשל היכולות והידע בשפה וגם בגלל פונקציונליות גבוהה. בשאר העבודה השתמשתי ב-Voyant ו-Dicta בשל היותם כלי מובנה לניתוח טקסטים ארוכים ללא שימוש במשאבי המחשב או יצירת כלי דומה מאפס.

לאחר קבלת תוצאות, נדרש לראות במה מתאפיין פירוש רש"י לעומת פירושים אחרים. לאחר מכן, להשוות כל פירוש במחלוקת לפירוש רש"י ולפירושים אחרים ולראות באילו מדדים הוא יותר קרוב או רחוק לרש"י.

ממצאים

1. זיהוי ואפיון סגנון רש״י לעומת אחרים באמצעות המדדים 1-7

הכנסתי את הטקסט הנקי של מסכתות שפירש רש״י ומסכתות שפירשו פרשנים אחרים לסקריפט עבור קביעת המאפיינים. כאמור, לקחתי מדגם רחב כל הניתן של מסכתות מסדר זרעים, נשים, קדשים, מועד ונזיקין באורכים ונושאים משתנים. המדגם הרחב נועד להבטיח תוצאות לא מוטות מאורך המסכת, נושא המסכת או מידת הקושי של המסכת.

לאחר עיבוד וחישובים סטטיסטים התוצאות בממוצע לפרשן הם:

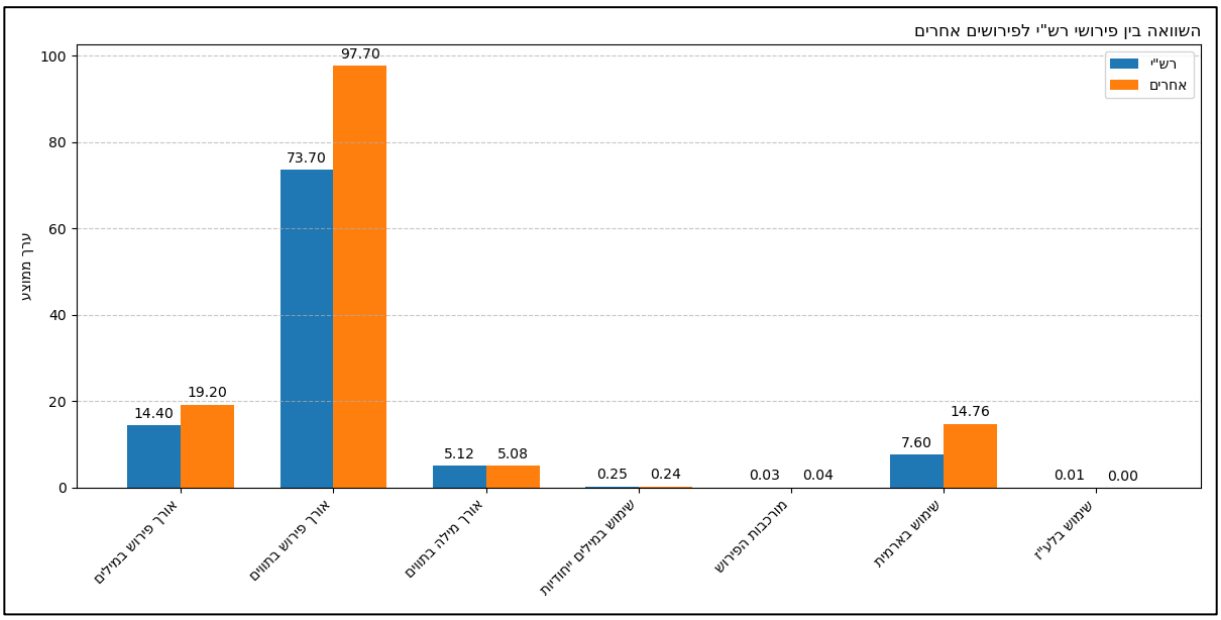
פרשן	**אורך פירוש במילים**	**אורך פירוש בתווים**	**אורך מילה בתווים**	**שימוש במילים יחודיות**	**מורכבות הפירוש**	**שימוש בארמית**	**שימוש בלע״ז**
רש״י	14.4	73.7	5.12	0.248	0.03	7.6	0.01
אחרים	19.2	97.7	5.08	0.244	0.04	14.76	0.01>

טבלה 2: תוצאות השוואת מדדים 1-7 בין רש״י לאחרים.

כשמנתחים את התוצאות ניתן לראות הבדל סגנוני ברור:

- פירוש רש״י מתאפיין באורך קצר יותר, גם במילים וגם בתווים. (14.4<19.2, 73.7<97.7)
- רש״י משתמש בפירושו ביותר מילים ייחודיות מפרשנים אחרים.
- פירוש רש״י פחות מורכב (0.03 לעומת 0.04 שימוש במילות קישור).
- רש״י משתמש בפירושו בפחות בארמית לעומת עברית (פי 2 פחות לעומת אחרים).
- בפירוש רש״י יש נוכחות גבוהה יותר של מילות לע״ז בפירוש.

ההבדלים בצורה גרפית:



איור 3 – גרף השוואה בין פירוש רש״י לאחרים.

לסיכום: פירוש רש"י תמציתי, קצר, פשוט ומתעדף עברית על פני ארמית.

2. **השוואת הסגנון למול הפירושים השנויים במחלוקת באמצעות מדדים 1-7**

א. <u>מסכת תענית</u>

לקחתי את מסכת תענית והעברתי אותה את אותו תהליך, לאחר מכן ערכתי השוואה למול התוצאות בסעיף הקודם:

פרשן	**אורך פירוש במילים**	**אורך פירוש בתווים**	**אורך מילה בתווים**	**שימוש במילים יחודיות**	**מורכבות הפירוש**	**שימוש בארמית**	**שימוש בלע"ז**
רש"י	14.4	73.7	5.12	0.248	0.03	7.6	0.01
אחרים	19.2	97.7	5.08	0.244	0.04	14.76	0.01>
תענית	**16.5**	**81.12**	**4.91**	**0.30**	**0.03**	**6.47**	**0.01**

טבלה 3: תוצאות השוואת המדדים 1-7 בין מסכת תענית לרש"י ואחרים.

כדי לקבל מסקנה נדרש להסתכל על המרחק בין הנתונים של תענית לבין רש"י ואחרים. מרחק קטן יותר יראה דמיון סגנוני לאחד מהם. חישבתי את המרחקים והגעתי למסקנות הבאות:

	אורך פירוש במילים	**אורך פירוש בתווים**	**אורך מילה בתווים**	**שימוש במילים יחודיות**	**מורכבות הפירוש**	**שימוש בארמית**	**שימוש בלע"ז**
תענית דומה ל-	רש"י	רש"י	אחרים	רש"י	רש"י	רש"י	רש"י

טבלה 4: דמיון המדדים של מסכת תענית לרש"י או אחרים.

ניתן לראות שהפירוש לתענית מתנהג כמו פירוש רש"י בכל הפרמטרים למעט "אורך מילה בתווים". כלומר, בממוצע הפרשן בתענית משתמש במילים קצרות יותר. **ניתן לטעון שאם מסתכלים רק על פרמטרים 1-7 רש"י אכן כתב את הפירוש למסכת תענית.**

ב. <u>מסכת מעילה</u>

לקחתי את מסכת מעילה והעברתי אותה את אותו תהליך, לאחר מכן ערכתי השוואה למול התוצאות בסעיף 1.:

פרשן	**אורך פירוש במילים**	**אורך פירוש בתווים**	**אורך מילה בתווים**	**שימוש במילים יחודיות**	**מורכבות הפירוש**	**שימוש בארמית**	**שימוש בלע"ז**
רש"י	14.4	73.7	5.12	0.248	0.03	7.6	0.01
אחרים	19.2	97.7	5.08	0.244	0.04	14.76	0.01>
מעילה	**23.67**	**122.3**	**5.16**	**0.25**	**0.05**	**2.82**	**0.01>**

טבלה 5 : תוצאות השוואת המדדים 1-7 בין מסכת מעילה לרש"י ואחרים.

כדי לקבל מסקנה נדרש להסתכל על המרחק בין הנתונים של מעילה לבין רש"י ואחרים. מרחק קטן יותר יראה דמיון סגנוני לאחד מהם. חישבתי את המרחקים והגעתי למסקנות הבאות:

	אורך פירוש במילים	**אורך פירוש בתווים**	**אורך מילה בתווים**	**שימוש במילים יחודיות**	**מורכבות הפירוש**	**שימוש בארמית**	**שימוש בלע"ז**
מעילה דומה ל-	אחרים	אחרים	רש"י	רש"י	אחרים	רש"י	אחרים

טבלה 6 : דמיון המדדים של מסכת מעילה לרש"י או אחרים.

ניתן לראות שהפירוש למסכת מעילה מתנהג ברוב הפרמטרים לפירושים אחרים, אך באופן פחות חד משמעי. ניתן לראות שהפירוש למסכת מעילה ארוך יותר בממוצע, מורכב יותר ואינו משתמש במילות לע"ז כמו פירוש רש"י המובהק. מאפיינים אלו הם בין הבולטים ביותר בפירוש רש"י. **לכן, ניתן לטעון שאם מסתכלים רק על פרמטרים 1-7 סביר יותר שרש"י <u>לא</u> כתב את הפירוש למסכת מעילה.** זאת כיוון שרוב המדדים מצביעים על קרבה למחברים אחרים.

ג. <u>מסכת הוריות</u>

לקחתי את מסכת הוריות והעברתי אותה את אותו תהליך, לאחר מכן ערכתי השוואה למול התוצאות בסעיף 1.:

פרשן	**אורך פירוש במילים**	**אורך פירוש בתווים**	**אורך מילה בתווים**	**שימוש במילים יחודיות**	**מורכבות הפירוש**	**שימוש בארמית**	**שימוש בלע"ז**
רש"י	14.4	73.7	5.12	0.248	0.03	7.6	0.01
אחרים	19.2	97.7	5.08	0.244	0.04	14.76	0.01>
הוריות	**19.5**	**98.2**	**5.04**	**0.24**	**0.03**	**0.85**	**0.01>**

טבלה 7 : תוצאות השוואת המדדים 1-7 בין מסכת הוריות לרש"י ואחרים.

כדי לקבל מסקנה נדרש להסתכל על המרחק בין הנתונים של הוריות לבין רש"י ואחרים. מרחק קטן יותר יראה דמיון סגנוני לאחד מהם. חישבתי את המרחקים והגעתי למסקנות הבאות:

	אורך פירוש במילים	**אורך פירוש בתווים**	**אורך מילה בתווים**	**שימוש במילים יחודיות**	**מורכבות הפירוש**	**שימוש בארמית**	**שימוש בלע"ז**
הוריות דומה ל-	אחרים	אחרים	אחרים	אחרים	רש"י	רש"י	אחרים

טבלה 8 : דמיון המדדים של מסכת הוריות לרש״י או אחרים.

ניתן לראות שהפירוש למסכת הוריות מתנהג ברוב הפרמטרים לפירושים אחרים. ניתן לראות שהפירוש למסכת הוריות ארוך יותר בממוצע, ואינו משתמש במילות לע״ז כמו פירוש רש״י המובהק. לעומת זאת, הפירוש אינו מורכב באופן יחסי ומעדיף שלא להשתמש בארמית. **ניתן לטעון שאם מסתכלים רק על פרמטרים 1-7 סביר יותר שרש״י לא כתב את הפירוש למסכת הוריות.** זאת כיוון שרוב המדדים מצביעים על קרבה למחברים אחרים.

3. **השוואת סגנונות למול מדד 8 - מילים נפוצות (טבלת שכיחויות)**

חישוב טבלת השכיחויות נעשה על כלל הפירושים (רש״י, אחרים, שנוי במחלוקת). חילצתי מהטקסט את המילים ללא כפילויות וספרתי כמה פעמים מופיעה מילה בכל טקסט. לאחר מכן, חישבתי את השכיחות של כל מילה בכל טקסט בנפרד. השכיחות חושבה לפי סה״כ מופעי המילה לחלק לסה״כ מופעי מילים בטקסט. כיוון שהשוואה היא למול רש״י, מיינתי את הרשימה לפי שכיחות המילים בטקסט של רש״י.

לאחר מכן מדדתי את המרחק בין השכיחות של כל מילה אצל הפירושים השנויים במחלוקת למול רש״י ואחרים ורשמתי את התוצאה בצורה של תג רש״י/אחרים/דומה.

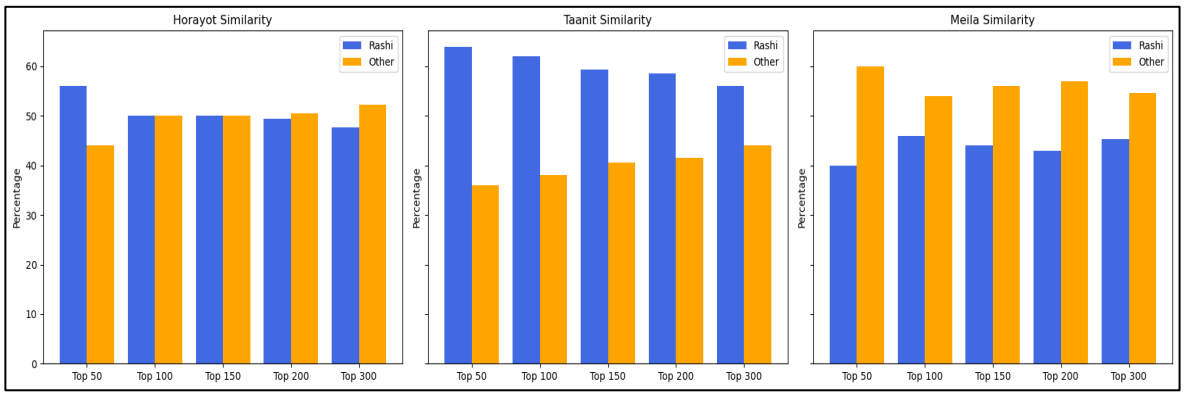
התקבלה טבלה שנראית ככה :

קרבה			שכיחות					מופעים					
מעילה	תענית	הוריות	מעילה	תענית	הוריות	אחרים	רש"י	מעילה	תענית	הוריות	אחרים	רש"י	מילה
אחרים	רש"י	רש"י	0.017	0.012	0.012	0.018	0.014	286	311	164	1752	12708	לא
אחרים	אחרים	רש"י	0.004	0.007	0.013	0.008	0.008	74	186	170	762	7506	על
אחרים	אחרים	רש"י	0.005	0.006	0.009	0.006	0.008	89	151	122	599	7387	הוא
אחרים	רש"י	רש"י	0.009	0.007	0.007	0.010	0.008	148	176	92	963	6862	אלא
רש"י	רש"י	רש"י	0.006	0.005	0.004	0.008	0.007	99	122	59	754	5945	ליה

טבלה 9 : טבלת שכיחויות מילים, 5 המילים הנפוצות ביותר בפירוש רש״י.

בנספח 5, ניתן לצפות בטבלה המלאה. כעת נותר לספור כמה פעמים כל מסכת קרובה לרש״י לעומת אחרים. בחנתי את מידת הקרבה לפי אינטרוולים של 50,100,150,200,300 המילים הנפוצות ביותר בפירוש רש״י. כך יהיה ניתן לבחון את היציבות של המדד למול מדגם הולך וגדל. למול 200 ומעלה מילים ייתכן שניתקל במילים שהם ערכי-קיצון (outliers) ולמול 50 המילים הכי נפוצות ייתכן שמרחב המדגם אינו מספיק רחב.

החישוב בוצע באמצעות סקריפט בשפת פייתון שסופר כמה פעם המיזוג הוא לרש״י לעומת סיווג לפרשנים אחרים באחוזים. הגרף להלן מראה את האחוזים מתוך המילים שסווגו כקרובות לרש״י למול קרבה לאחרים, באינטרוולים של 50,100,150,200,300 המילים הנפוצות ביותר :



איור 4: תרשים עמודות קרבה לרש"י למול קרבה לאחרים, לפי כמות מילים נפוצות.

מימין לשמאל: מסכת מעילה, מסכת תענית ומסכת הוריות. העמודות מייצגות באחוזים קרבה לרש"י או לאחרים. רש"י מיוצג בצבע כחול, אחרים מיוצג בצבע כתום.

<u>התוצאות</u> (לפי המסכתות בגרף מימין לשמאל):

א. <u>מסכת מעילה</u>

מהתוצאות מתקבל ששכיחויות המילים הם באופן עקבי דומה יותר לפרשנים אחרים מאשר לפירוש רש"י. הדמיון לפרשנים אחרים אף מתחזק בין 100 מילים ל-200 מילים ולאחר מכן נשאר עקבי. ניתן להכריע באופן ברור לפי מדד זה **שהפירוש למסכת מעילה אינו נכתב ע"י רש"י.**

ב. <u>מסכת תענית</u>

מהתוצאות מתקבלת קרבה לפירוש רש"י באופן ברור עקבי. ניתן להכריע באופן ברור לפי מדד זה **שהפירוש למסכת תענית נכתב ע"י רש"י**.

ג. <u>מסכת הוריות</u>

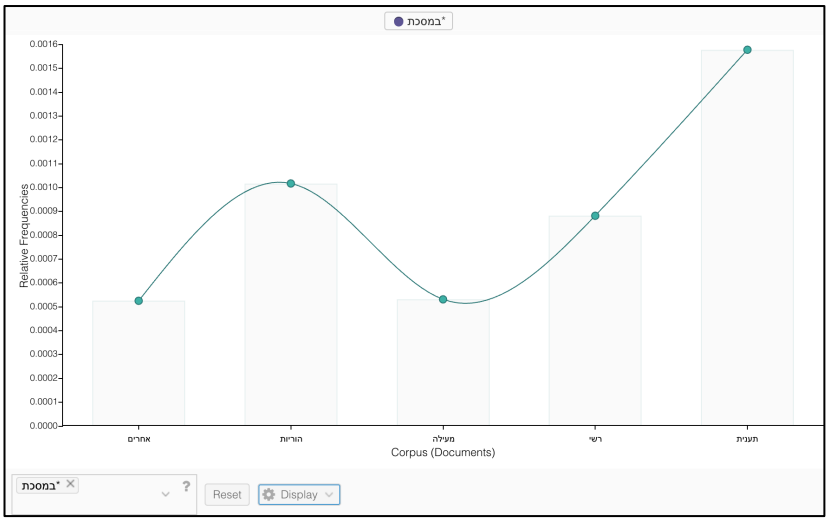
החל מהשוואה ל-100 מילים ועד ל-200 מילים התוצאות אינן חד משמעיות ונעות בין קרבה לפירוש רש"י לבין קרבה לפרשנים אחרים, אך גם קרבה זו שולית ונמצאת בטווח הטעות הסטטיסטית. ולכן **לא ניתן להכריע לפי מדד זה**.

4. זיהוי ואפיון סגנון באמצעות Voyant

Voyant הוא אתר חינמי המאפשר ניתוח של טקסטים מרובים באמצעות שיטות שונות של מדעי הרוח הדיגיטליים. באמצעות האתר ניתן להעלות טקסטים בשפות שונות וקבל ניתוח כמו שכיחות מילים, קשרים בין מילים ומגמות בטקסט גם תוך כדי תצוגה וויזואלית עם גרפים מסוגים שונים. הכלים באתר משמשים לניתוח וויזואלי אינטראקטיבי של טקסטים ללא מגבלת גודל הטקסט.

למרות שאין הגבלה על גודל הטקסט, פתיחה של טקסטים גדולים מכבידה על האתר והעיבוד הוויזואלי. לכן, לקחתי מכל מסכת 10,000 תווים ראשונים ויצרתי 5 סוגי טקסט: רש"י, אחרים, תענית, הוריות ומעילה. הזנתי את הקבצים באתר והתחלתי לאסוף תובנות.

א. <u>הפנייה למסכת אחרת</u>



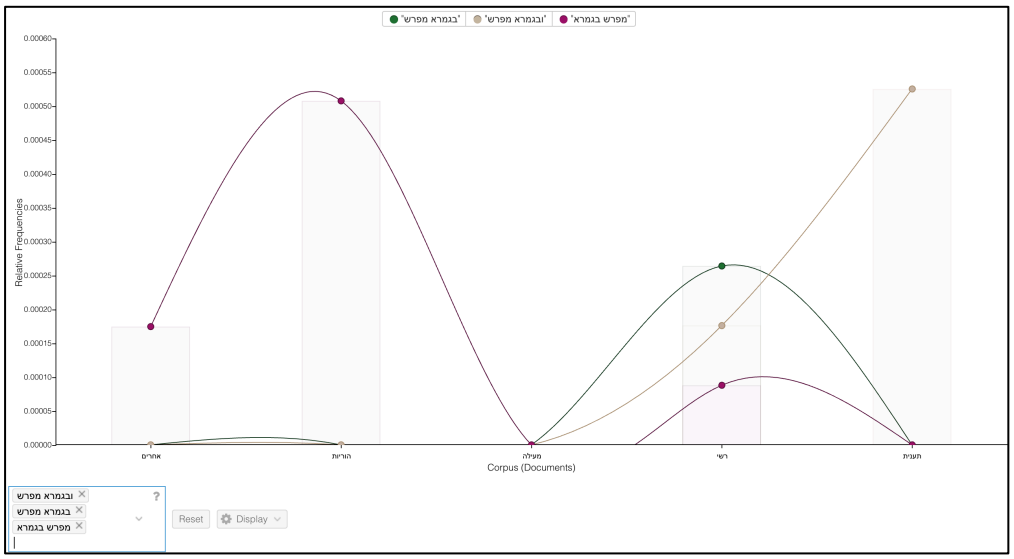
איור 5: התפלגות מופעי המילה "במסכת" המעידה על הפנייה למסכת אחרת.

באיור 5 ניתן לראות באיזה קובץ מופיעה ההפניה למסכת אחרת ובאיזו תדירות.

- ניתן לראות הבדל ברור בין רש"י ל-אחרים.
- קיים דמיון בין רש"י להוריות ובין אחרים למעילה.
- התדירות במסכת תענית קרובה יותר לרש"י מלאחרים.

ב. <u>הפנייה או הזכרת הפירוש בגמרא</u>

ישנם כמה ביטויים שבהם המפרש משתמש כדי להזכיר או להפנות לפירוש הגמרא לעניין מסוים. אפילו שהביטויים דומים הם מעידים על שוני סגנוני.



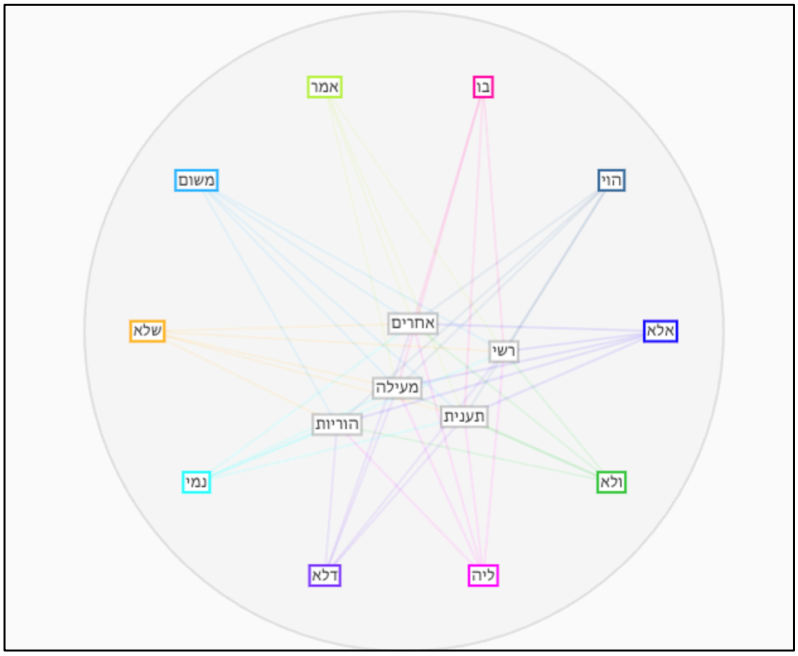
איור 6: התפלגות השימוש בביטוי "מפרש בגמרא" והטיותיו השונות בסוגי הטקסט.

- באיור 6 ניתן לראות שרש"י משתמש בכל שלושת הביטויים, אבל אחרים משתמשים רק ב-"מפרש בגמרא".
- בתענית נעשה שימוש רב ב-"ובגמרא מפרש", ביטוי שאחרים כמעט לא משתמשים בו אבל רש"י כן משתמש בו.
- במסכת מעילה אין כמעט אף ביטוי דומה.

ג. <u>קשרים בין ביטויים לטקסט</u>

אחד הגרפים שניתן להציג באתר נקרא Mandala. גרף זה מראה קשרים בין ביטויים וטקסטים. כל ביטוי "מושך" אליו טקסט לפי התדירות היחסית שלו, כך שביטוי יהיה קרוב יותר לטקסט אם הוא מופיע בתדירות גבוהה בו, ולהפך.

הגרף גם יוצר קיבוץ אשכולות של הטקסטים (Clustering) כך שטקסטים שצמודים זה לזה יותר דומים ולהפך.



איור 7: גרף Mandala המראה קשרים בין ביטויים למסמכים.

באיור 7 ניתן לראות את המרחקים בדמיון בין הטקסטים כתוצאה מהתדירות של המילים הנפוצות ביותר. לפי גרף זה ניתן להסיק כי:

- פירוש רש"י קרוב לנוס הפירוש במסכת תענית, ביחס לשאר הפירושים.
- פירוש אחרים קרוב לנוסח הפירוש במסכת מעילה, יותר מלכל טקסט אחר.
- מסכת הוריות נראית מרוחקת מאוד ביחס לכל הטקסטים.

ד. <u>מסקנות</u>:

הודות לניתוחים הוויזואליים שסיפק אתר Voyant ניתן להגיע לתובנות הדומות לתובנות מן הניתוחים הסטטיסטיים:

I. קיים שוני סגנוני בין פירוש רש"י לפירוש אחרים. רש"י ואחרים משתמשים בביטויים שונים ובתדירות שונה וקיים מרחק ניכר ביניהם בגרף ה-Mandala.

II. פירוש רש"י ונוסח הפירוש בתענית קרובים זה לזה בשימוש בביטויים דומים ובתדירות המילים הנפוצות. קיימת קרבה יחסית בין פירוש רש"י לפירוש במסכת תענית.

5. סיכום הממצאים

לסיכום, ניתן לראות שהפירוש למסכת תענית נכתב על ידי רש"י, בעוד הפירוש למסכת הוריות והפירוש למסכת מעילה ככל הנראה נכתבו על ידי פרשנים אחרים. יש לציין שבדומה למסכתות אחרות (שלא הובאו במאמר זה) קיים סיכוי שהפירוש למסכתות הוריות ומעילה מעורב בפירוש רש"י שלא השתמר במלואו ולכן נראית התאמה חלקית במדדים מסוימים. בנוסף, ראוי לציין שחסרון של מסכת מסדר קדשים ייתכן והשפיע על התוצאות של מסכת מעילה.

בטבלה מטה ניתן לראות סיכום התוצאות של כל המדדים :

מדד סגנוני	מסכת תענית	מסכת הוריות	מסכת מעילה
אורך פירוש (במילים)	דומה לרש"י	דומה לאחרים	דומה לאחרים
אורך פירוש (בתווים)	דומה לרש"י	דומה לאחרים	דומה לאחרים
אורך מילה ממוצע	דומה לאחרים	דומה לאחרים	דומה לרש"י
שימוש במילים ייחודיות	דומה לרש"י	דומה לאחרים	דומה לרש"י
מורכבות הפירוש (שימוש במילות קישור)	דומה לרש"י	דומה לרש"י	דומה לאחרים
שימוש בארמית	דומה לרש"י	דומה לרש"י	דומה לאחרים
שימוש בלע"ז	דומה לרש"י	דומה לאחרים	דומה לאחרים
שכיחות מילים (מדד טבלת שכיחויות)	דומה לרש"י	לא חד משמעי	דומה לאחרים
ניתוח Voyant (ביטויים, Mandala)	דומה לרש"י	דומה לאחרים	דומה לאחרים

טבלה 10 : סיכום תוצאות הממדים.

דיון ומסקנות

הממצאים שהתגלו במחקר זה תומכים בדעות של גדולי ישראל וחוקרים מודרניים כאחד, הן באשר לאפיון הסגנוני של פירוש רש״י והן בהכרעה בשאלת הייחוס של הפירוש לתענית, מעילה והוריות. בניגוד לניתוח לשוני ועיוני מסורתי בפירושים אלו שנעשה בעבר, מחקר זה השתמש בכלי ניתוח סגנוני ממחושבים כדי לבחון את שאלות המחקר באמצעיים מודרניים.

התמיכה של הממצאים בייחוס הפירוש למסכת תענית לרש״י תואמת לשיטתם של החיד״א, הריטב״א ועוד, אך עומד בניגוד לדעתו של הרב חיות. ממצאים אלו מראים כיצד מאפיינים כמו אורך פירוש, מורכבות הפירוש והעדפת השפה העברית מהווים חתימה סגנונית לפירושו של רש״י. לעומת זאת, הפירוש למסכמת מעילה ולמסכת הוריות מתרחקים ברוב המדדים מהסגנון של פירוש רש״י. ממצאים אלו תואמים לדעת הרב מלאכי הכהן ופרופ׳ אפשטיין שפירושים אילו לא נכתבו על ידי רש״י.

ההשלכות של מחקר זה הן מעבר למענה על שאלות המחקר. מחקר זה מוכיח את הכוח שיש לניתוח סגנוני בעולם מדעי הרוח הדיגיטליים בזיהוי מחברים גם לטקסטים עתיקים, מסורתיים ובשפה הארמית. בכך נפתחת הדרך למחקרים נוספים שיעסקו בחקר ייחוס או זיוף לכתבים אחרים בעולם מדעי הרוח, ובפרט במחברים מן הספרייה היהודית.

בנוסף, שיטות ניתוח כמו אלו במחקר זה אינם מצריכות הבנה בנושא הטקסט, אלא נשענות על מאפיינים סגנוניים בלבד. יש בעובדה זו כדי לחזק את הממצאים, אך גם יש לבצע עבודת ניקוי ואימות לטקסטים כדי להיזהר מהטיות.

סיכום

המחקר מראה כיצד שיטות של מדעי הרוח הדיגיטלי מיושמים כדי לזהות ולאפיין סגנון של כותב ואף לסווג טקסטים אחרים ביחס קרבה או ריחוק אליו. באמצעות שימוש במאפיינים סטטיסטים כמו אורך הפירוש, שכיחות מילים ועוד ניתן לקבל את טביעת האצבע הסגנונית של הכותב ולהבחין בינה לבין סגנון של כתובים אחרים. השיטה שננקטה כאן ממחישה כיצד ניתן לשלב בין כלים חדשניים לטקסטים תורניים עתיקים בכדי לענות של שאלות של ייחוס כתבים במחלוקת. בנוסף, מחקר זה פותח דלת להמשך זיהוי וחקירה של טקסטים תלמודיים ורבניים אחרים בשביל להעמיק את היכרותנו עם מחברי הטקסטים החשובים לספרייה היהודית.

ביבליוגרפיה:

1. אזולאי, ח. י. ד. (1998). שם הגדולים: מוסד הרב קוק. (מהדורת צילום ממהדורת ליוורנו, תקמ"ו)

2. אפשטיין, י. נ. (1995). המיוחס לרש"י בהוריות. תרביץ, 64(4), 587–601.

3. דבליצקי, ב. (2015). מקדש בציון (מהדורת סריקה). בני ברק: הוצאת יד מהרי"ץ.

4. וינטר, י., & קופל, מ. (2014). Determining if two documents are written by the same author. arXiv. https://arxiv.org/abs/1404.3185

5. חיות, צ. ה. (1959). אמרי בינה. ירושלים: מוסד הרב קוק.

6. כהן, מ. (1959). יד מלאכי. ירושלים: מכון ירושלים. (מהדורת צילום מדפוס וילנא, תקנ"ה)

7. קופל, מ. (2010). זיהוי מחברים בשיטות ממוחשבות: המקרה של גניזת חרסון. ישורון, כ"ב (אלול תש"ע), 837–849.

8. קופל, מ. (2011). Digital Stylistics for Authorship Attribution. הרצאה ב־Bar-Ilan NLP Seminar. https://u.cs.biu.ac.il/~koppel/papers/digital_stylometry.pdf

9. קנפנטון, י. (2006). דרכי התלמוד (הקדמה מאת י. קאפח). ירושלים: מוסד הרב קוק.

10. פרנקל, י. (2001). דרכו של רש"י בפירושו לתלמוד הבבלי. רמת גן: אוניברסיטת בר־אילן.

11. רבי יצחק די לאטיש. (1995). בתוך: גרוסמן, א. (עורך), תורת הצפון ג' (עמ' 155–157). ירושלים: מוסד הרב קוק.

נספחים:

נספח 1 – מדד העדפת ארמית

המדד עוזר לקבוע האם הפרשן מעדיף ארמית על פני עברית. דרך החישוב הייתה לספור כמה פעמים פרשן משתמש בכל מילה, ואז לחלק את מספר המופעים של ארמית במספר המופעים של עברית. נבחרו מילים בשימוש גבוה בגמרא, כי דווקא מהמילים שמשתמשים בהם באופן שוטף אפשר לראות את הסגנון הייחודי של הפרשן.

המילים שנבחרו:

עברית	שלא	גם	שאמר	תאמר	שם	מה	לו	שנינו
ארמית	דלא	נמי	דאמר	תימא	התם	מאי	ליה	תנן

טבלה 11: מילים בערבית וארמית.

נספח 2 – מדד מורכבות הפירוש

מדד זה עוזר לקבוע כמה הפירוש מורכב, ככל שיש שימוש רב יותר במילות קישור, כך הפרשן מתקשה בהסבר פשט הכתוב.

מילות הקישור שנבחרו וביטויים מקשרים:

מילות קישור			
אמאי	ועוד	אלא	דא
בעי	דכוותה	ואי	דהא
איבעיא	כגון	ואי נמי	הואיל
לעולם	הכי	ואם	היכי
אדרבה	הכי נמי	ואם תמצא לומר	הכא
מיהו	ולאו	אלא	התם
משום	ולא	אי נמי	השתא
כיון	מאי	ואי תימא	עד
כל שכן	למה	אפילו	אי
ואם	ויש		

טבלה12: מילות קישור וביטויים דומים.

נספח 3 – קוד

א. קוד שיוצר את טבלת הסטטיסטיקות של המדדים ומחשב את המרחקים בין המסכתות לבין רש״י ואחרים:

```
1. """
2. Iterate over source files and make data table of statistics
3. """
4.
5. import os
6. import pandas as pd
7. from collections import defaultdict
8. from typing import List, Dict, Any, Tuple
9. import string
10.
11. # List of Hebrew stopwords
12. hebrew_stopwords = set([
13.     'של', 'על', 'עם', 'את', 'זה', 'היא', 'הוא', 'הם', 'הן', 'אני', 'אתה', 'את',
    'אנחנו', 'הוא', 'היא', 'הם', 'הן',
14.     'מה', 'מי', 'זה', 'זאת', 'אלה', 'אלו', 'כן', 'לא', 'אם', 'או', 'אבל', 'כי',
    'כאשר', 'אשר', 'אף', 'כל', 'רק',
15.     'כמו', 'למה', 'לכן', 'למרות', 'אחרי', 'לפני', 'בין', 'תחת', 'מאחורי', 'מעל',
    'מתחת', 'מול', 'ליד', 'אצל',
16.     'ב', 'ל', 'מ', 'כ', 'ש', 'ו', 'ה', 'י', 'ת', '–', '—'
17. ])
18. # Statistic data table columns
19. COLUMNS = [
20.     'file',
21.     'num_pages',
22.     #'num_refs',
23.     #'common_words',
24.     'avg_comment_words',
25.     'avg_comment_char',
26.     'avg_comment_char_per_word',
27.     'max_comment_len_words',
28.     'min_comment_len_words',
29.     'loazi_words_count',
30.     'avg_comments_wLoazi',
31.     'unique_words_count',
32.     'total_words_count',
33.     'unique_words_usage_rate',
34.     'comment_complexity',
35.     'masechet']
36. stt_df = pd.DataFrame(columns=COLUMNS)
37. comment_lengths_df = pd.DataFrame(columns=['masechet', 'length'])
38. word_count = defaultdict(int)
39.
40. # Dictionary for Talmud masechet to Rashi or other
41. talmud_dict = {
42.     "megila": "rashi",
43.     "sanhendrin": "rashi",
```

```
    "brachot": "rashi",
    "psahim": "rashi",
    "beitza": "rashi",
    "sukka": "rashi",
    "nazir": "other",
    "horayot": "unknown",
    "shabat": "rashi",
    "taanit": "unknown",
    "hagiga": "rashi",
    "eruvin": "rashi",
    "nedarim": "other",
    "moed": "mixed",
    "pasahimRashbam": "other",
    "meila": "unknown",
    "yoma": "rashi",
    "kiddushin": "rashi",
    "zevahim": "rashi",
    "babaMetsia" : "rashi",
    "babaKama": "rashi",
    "gittin" : "rashi"
}

def make_timestamp() -> str:
    " Make a timestamp for the output file "
    from datetime import datetime
    now = datetime.now()
    return now.strftime("%Y-%m-%d_%H-%M-%S")

# preprocess file
def preprocess_file(peirush: pd.DataFrame) -> pd.DataFrame:
    " Preprocess a file for analysis "
    # remove empty rows
    peirush = peirush.dropna()
    # replace all the "–" with "-" from
    peirush.iloc[:,0] = peirush.iloc[:,0].replace("–", "-")
    # Remove parentheses and content within them
    # add a "#" to mark the removed content for future reference
    peirush.iloc[:,0] = peirush.iloc[:,0].str.replace(r"\(.*\)", "#", regex=True)

    return peirush

def aramit_preference_rate(text: List[str]) -> float:
    "Calculate the rate of Aramaic words usage in relation to Hebrew words"
    heb_arm = {
        'שלא': 'דלא',
        'גם': 'נמי',
        'שאמר': 'דאמר',
        'תאמר' : 'תימא',
        'שם': 'התם',
```

```
        'מה': 'מאי',
        'לו': 'ליה',
        'שנינו': 'תנן'
    }
    rates = []
    for heb_word, arm_word in heb_arm.items():
        heb_count = text.count(heb_word)
            arm_count = text.count(arm_word)
            if heb_count > 0:
                rates.append(arm_count / heb_count)
            else: rates.append(1.0)

        # Words ending with "א" are Aramaic
        aramaic_count = sum(1 for word in text if word.endswith('א'))
        # Words ending with "ה" are Hebrew
        hebrew_count = sum(1 for word in text if word.endswith('ה'))
        if hebrew_count > 0:
            rates.append(aramaic_count / hebrew_count)

        return sum(rates) / len(rates) if rates else 0.0

    def comment_complexity(text: List[str]) -> Tuple[float, float]:
        "Calculate the average word length and the average sentence length"
        linking_words = [ "דא",  "דהא", "הואיל", "היכי", "הכא", "התם", "השתא", "עד", "אי", "אלא", "ואי", "ואי נמי", "ואם", "ואם תמצא לומר", "ואם תמצא לומר", "אלא",  "אי נמי",
                          "ואי תימא",
                          "אפילו", "ועוד", "דכוותה", "כגון", "הכי", "הכי נמי", "אי נמי", "ולאו", "ולא", "מאי", "למה",
                            "אמאי", "בעי", "איבעיא", "לעולם", "אדרבה", "מיהו", "משום", "כיון", "כל שכן", "ואם", "ויש"]
        # count linking words in text
        text_complexity = 0
        for comment in text:
            if len(comment) == 0: continue
            text_complexity += sum(1 for word in comment if word in linking_words) / len(comment)

        return text_complexity / len(text) if len(text)>0 else 0.0

    def format_df(df: pd.DataFrame) -> pd.DataFrame:
        " Format the data frame to perush and dibur"
        formatted_df = pd.DataFrame(columns=['dibur', 'comment', 'page'])
        now_page = ''

        for _,content in df.iterrows():
            print(f"\r{df.columns[0]} | {now_page} | {content.iloc[0]}", end='', flush=True)
            if content.iloc[0].startswith("Daf"):
```

```
                now_page = content.iloc[0].split(' ')[1]

            elif not content.iloc[0].startswith("Line"):
                dibur_rows = content.iloc[0].split(': ')
                for dibur_row in dibur_rows:
                    split_content = dibur_row.split('-',1)
                    if len(split_content) == 1:
                        split_content.insert(0, None)

                    dibur, comment = split_content
                    formatted_df = formatted_df._append({'dibur': dibur,
                                                         'comment': comment,
                                                         'page': now_page},
                                                        ignore_index=True)
        return formatted_df

    def common_words_stats(text: List[str]) -> Dict[str, int]:
        " Calculate the common words in the text "
        global word_count
        # Count the occurrences of each word
        for word in text:
            if word not in hebrew_stopwords:
                word_count[word] += 1
        return dict(word_count)

    def get_file_stats(file_path: str) -> Dict[str, Any]:
        """Get statistics for a single file."""
        global comment_lengths_df
        try:
            with open(file_path, 'r', encoding='utf-8') as file:
                print(f"Reading file: {file_path}")
                new_row = {'file': file_path}

                # Load Excel file into a DataFrame
                subdata = pd.read_excel(file_path)
                subdata = subdata.dropna()  # Remove empty rows

                # Remove parentheses and content within them
                subdata.iloc[:, 0] = subdata.iloc[:, 0].str.replace(r"\(.*\)", "", regex=True)

                # Format the DataFrame
                formatted_df = format_df(subdata)

                # Remove punctuation from text
                translator = str.maketrans('', '', string.punctuation + '-:')
```

```
                formatted_df['dibur'] = formatted_df['dibur'].str.translate(translator)
                formatted_df['comment'] = formatted_df['comment'].str.translate(translator)

                # Extract "masechet" from the file name
                new_row['masechet'] = "".join(file_path.split('/')[-1].split('_')[-1]).replace('.xlsx', '')
                tag = talmud_dict.get(new_row['masechet'], 'unknown')

                # Number of pages
                new_row['num_pages'] = formatted_df['page'].nunique()

                # Average comment by words
                comment_lengths_words = formatted_df['comment'].str.split().str.len()
                new_row['avg_comment_words'] = comment_lengths_words.mean()

                # add comment_lengths_words to comment_lengths_df
                comment_lengths_words_df = pd.DataFrame(comment_lengths_words)
                comment_lengths_words_df['masechet'] = new_row['masechet']
                comment_lengths_df = comment_lengths_df._append(comment_lengths_words_df, ignore_index=True)

                # Average comment len by characters
                comment_lengths_chars = formatted_df['comment'].str.len()
                new_row['avg_comment_char'] = comment_lengths_chars.mean()

                # Average characters per word
                avg_chars_per_word = (
                    formatted_df['comment'].str.len().sum() /
                    formatted_df['comment'].str.split().str.len().sum()
                ) if formatted_df['comment'].str.split().str.len().sum() > 0 else 0
                new_row['avg_comment_char_per_word'] = avg_chars_per_word

                # Average linking words per sentence (text complexity)
                new_row['comment_complexity'] = comment_complexity(formatted_df['comment'].str.split())

                # Minimum comment length in words (ignoring empty comments)
                min_comment_len_words = comment_lengths_words.replace(0, float('inf')).min()
```

```
                    new_row['min_comment_len_words'] = 0 if min_comment_len_words == float('inf') else min_comment_len_words

                    # Maximum comment length in words
                    new_row['max_comment_len_words'] = comment_lengths_words.max()

                    # Count occurrences of "לע"ז"
                    new_row['loazi_words_count'] = formatted_df['comment'].str.contains('בלעז').sum()

                    # amount comments
                    num_comments = formatted_df['comment'].count()
                    # amount comments with "לע"ז"
                    num_loazi_comments = formatted_df['comment'].str.contains('בלעז').sum()

                    # Average "לע"ז" occurrences per comments
                    new_row['avg_comments_wLoazi'] = num_loazi_comments / num_comments

                    # Unique word count
                    all_words = ' '.join(formatted_df['comment']).split()
                    new_row['unique_words_count'] = len(set(all_words))

                    # total word count
                    new_row['total_words_count'] = len(all_words)

                    # unique words usage rate
                    new_row['unique_words_usage_rate'] = new_row['unique_words_count'] / new_row['total_words_count']

                    # Get aramit preference rate (masechet)
                    new_row['aramit_preference_rate'] = aramit_preference_rate(all_words)

                    print(f"{file_path} added to data table")
                return new_row

            except Exception as e:
                print(f"Error processing file {file_path}: {e}")
                raise e
                return {}

        def get_proximity(stt_df):
            # חישוב ממוצעים של כל תגית
            grouped_means = stt_df[stt_df['tag'] != 'unknown'].groupby("tag").mean(numeric_only=True)
            params = stt_df.drop(columns = ['tag', 'masechet']).columns
            proximity_df = pd.DataFrame(columns=params, index = stt_df[stt_df['tag'] == 'unknown']['masechet'])
```

```
258.         distace_df = pd.DataFrame(columns=params, index =
   stt_df[stt_df['tag'] == 'unknown']['masechet'])
259.
260.         for masechet in proximity_df.index:
261.             for param in params:
262.                 # Check if the parameter is numeric
263.                 if pd.api.types.is_numeric_dtype(stt_df[param]):
264.                     # Calculate the proximity to rashi or other for the
   masechet
265.                     val = stt_df.loc[stt_df['masechet'] == masechet,
   param].values[0]
266.                     distance_rashi = abs(val -
   grouped_means.loc['rashi', param])
267.                     distance_other = abs(val -
   grouped_means.loc['other', param])
268.                     # Calculate the distance to the average of the
   other tag
269.                     distace_df.loc[masechet, param] =
   min(distance_rashi, distance_other)
270.                     # Determine the proximity
271.                     if distance_rashi < distance_other:
272.                         proximity_df.loc[masechet, param] = 'rashi'
273.                     elif distance_rashi > distance_other:
274.                         proximity_df.loc[masechet, param] = 'other'
275.                     else:
276.                         proximity_df.loc[masechet, param] = 'equal'
277.
278.         return proximity_df, distace_df
279.
280.     if __name__ == '__main__':
281.         source_dir = r'/Users/nird/Library/CloudStorage/OneDrive-
   Personal/Uni/שנה ב/מדעי הרוח הדיגיטליים/rashi/sources'
282.         output_file = f'outputs/data_table_{make_timestamp()}.xlsx'
283.
284.         # Iterate over files in source directory
285.         for root, _, files in os.walk(source_dir):
286.             for file in files:
287.                 if file.endswith('.xlsx'):
288.                     file_path = os.path.join(root, file)
289.                     stt_df = stt_df._append(get_file_stats(file_path),
   ignore_index=True)
290.
291.         stt_df['masechet'] = stt_df.masechet.str.replace('.xlsx', '')
292.         stt_df['tag'] = stt_df['masechet'].map(talmud_dict)
293.         stt_df = stt_df[['tag',
294.                          'masechet',
295.                         'avg_comment_words',
296.                          'avg_comment_char',
297.                          'avg_comment_char_per_word',
298.                          'avg_comments_wLoazi',
299.                          'unique_words_usage_rate',
```

```
                                'comment_complexity',
                                'aramit_preference_rate']]

            comment_lengths_df['masechet'] = comment_lengths_df.masechet.str.replace('.xlsx', '')
            comment_lengths_df['tag'] = comment_lengths_df['masechet'].map(talmud_dict)

            proximity_df, distance_df = get_proximity(stt_df)

            with pd.ExcelWriter(output_file, engine='openpyxl', mode='w') as writer:
                comment_lengths_df.to_excel(writer, sheet_name='comment_lengths_df', index=False)
                stt_df.to_excel(writer, sheet_name='data_table', index=False)
                proximity_df.to_excel(writer, sheet_name='proximity_df', index=True)
                distance_df.to_excel(writer, sheet_name='proximity_df', index=True, startrow=proximity_df.shape[0] + 2)
            print('Data table written to', output_file)
```

ב. סקריפט פשוט ליצירת הגרף שמשווה בין רש״י לאחרים במדדים הסטטיסטים:

```
import matplotlib.pyplot as plt
import numpy as np

# פונקציה להפיכת טקסטים
def reverse_hebrew(texts):
    return [text[::-1] for text in texts]

# נתונים
labels = ['אורך פירוש במילים', 'אורך פירוש בתווים', 'אורך מילה בתווים',
          'שימוש במילים ייחודיות', 'מורכבות הפירוש', 'שימוש בארמית', 'שימוש בלע"ז']
rashi = [14.4, 73.7, 5.12, 0.248, 0.03, 7.6, 0.01]
others = [19.2, 97.7, 5.08, 0.244, 0.04, 14.76, 0.0005]

# הפיכת הטקסטים
labels_reversed = reverse_hebrew(labels)
title = 'השוואה בין פירושי רש"י לפירושים אחרים'[::-1]
ylabel = 'ערך ממוצע'[::-1]

x = np.arange(len(labels))
width = 0.35

fig, ax = plt.subplots(figsize=(12, 6))
bars1 = ax.bar(x - width/2, rashi, width, label='רש"י'[::-1])
```

```
bars2 = ax.bar(x + width/2, others, width, label='אחרים'[::-1])

ax.set_ylabel(ylabel)
ax.set_title(title, horizontalalignment='right', loc='right')
ax.set_xticks(x)
ax.set_xticklabels(labels_reversed, rotation=45, ha='right')
ax.legend()

def add_labels(bars):
    for bar in bars:
        height = bar.get_height()
        ax.annotate(f'{height:.2f}',
                    xy=(bar.get_x() + bar.get_width() / 2, height),
                    xytext=(0, 3),
                    textcoords="offset points",
                    ha='center', va='bottom')

add_labels(bars1)
add_labels(bars2)

plt.grid(axis='y', linestyle='--', alpha=0.7)
plt.tight_layout()
plt.show()
```

ג. קוד לחילוץ טקסט לפני עיבוד ב- Voyant

```
" extract heb text for voyant "
import os
import pandas as pd
from collections import defaultdict
from typing import List, Dict, Any, Tuple
import string
import project_make_data12 as pr_scripts

# Dictionary for Talmud masechet to Rashi or other
talmud_dict = {
    "megila": "rashi",
    "sanhendrin": "rashi",
    "brachot": "rashi",
    "psahim": "rashi",
    "beitza": "rashi",
    "sukka": "rashi",
    "nazir": "other",
    "horayot": "horayot",
    "shabat": "rashi",
    "taanit": "taanit",
    "hagiga": "rashi",
    "eruvin": "rashi",
    "nedarim": "other",
    "pasahimRashbam": "other",
    "meila": "meila",
```

```
    "yoma": "rashi",
    "kiddushin": "rashi",
    "zevahim": "rashi",
        "babaMetsia" : "rashi",
    "babaKama": "rashi",
    "gittin" : "rashi"
}

tag_text_dict = {
    'rashi': "",
    'other': "",
    'horayot': "",
    'taanit': "",
    'meila': ""
}

def make_timestamp() -> str:
    """Generate a timestamp string."""
    from datetime import datetime
    return datetime.now().strftime('%Y%m%d_%H%M%S')

def extract_heb_text(file_path):
    """Count the frequency of each word in a file."""
    global talmud_dict
    global tag_text_dict
    masechet = "".join(file_path.split('/')[-1].split('_')[-1]).replace('.xlsx',
'')
    tag = talmud_dict.get(masechet, 'unknown')

    # Load Excel file into a DataFrame
    subdata = pd.read_excel(file_path)
    subdata = subdata.dropna()  # Remove empty rows

    # Remove parentheses and content within them
    subdata.iloc[:, 0] = subdata.iloc[:, 0].str.replace(r"\(.*\)", "",
regex=True)

    # Format the DataFrame
    formatted_df = pr_scripts.format_df(subdata)

    # # Remove punctuation from text
    # translator = str.maketrans('', '', string.punctuation + '-:')
    # formatted_df['comment'] = formatted_df['comment'].str.translate(translator)

    formatted_df = formatted_df['comment']
    # remove non-Hebrew characters besides punctuation and spaces
    formatted_df = formatted_df.str.replace(r"[^א-ת\s.,:;?!]", "", regex=True)  #
Keep Hebrew letters and spaces
    #formatted_df = formatted_df.str.replace(r"\s+", " ", regex=True)  # Remove
extra spaces
    #formatted_df = formatted_df[formatted_df != ""]  # Remove empty strings
```

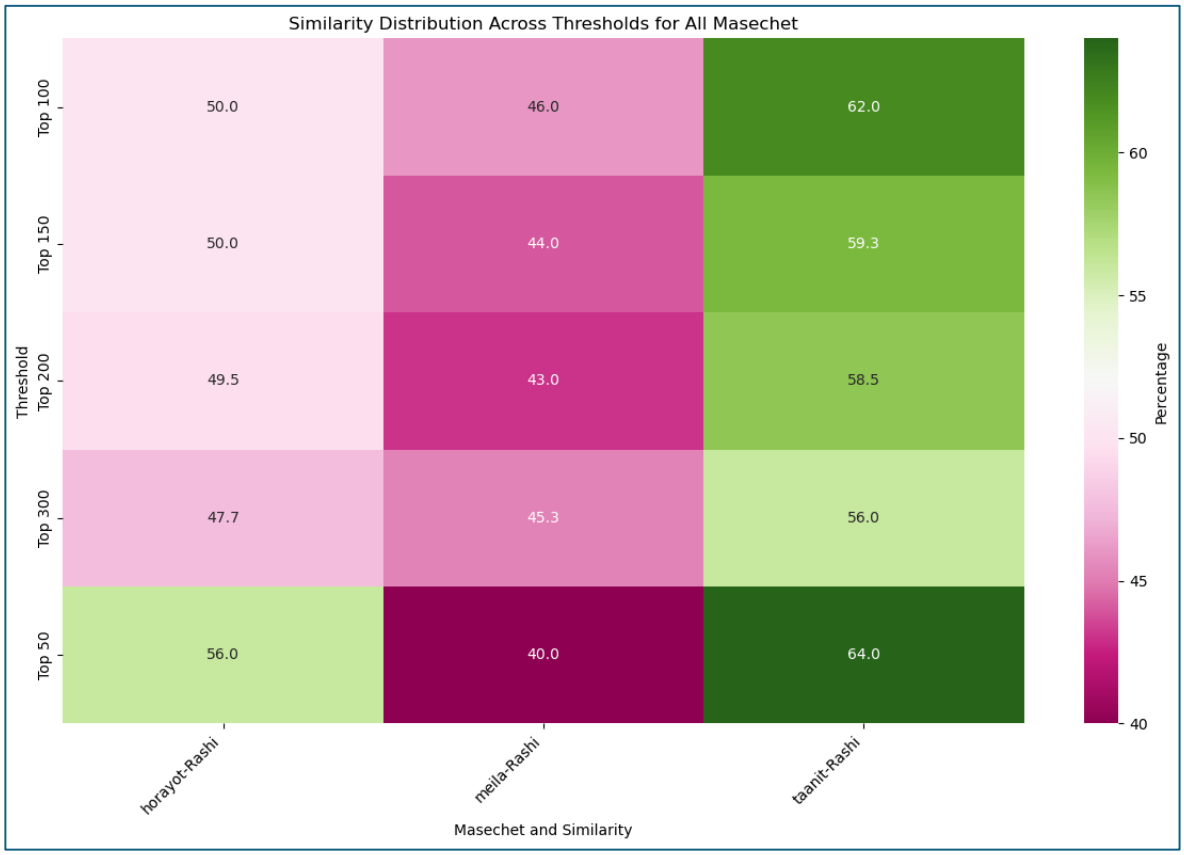
```
    # Convert to a single string
    text = " ".join(formatted_df.tolist())
    text = text.replace('\n', ' ')  # Replace newlines with spaces
    text = text.replace('\r', ' ')  # Replace carriage returns with spaces
    text = text.replace('\t', ' ')  # Replace tabs with spaces

    if len(text) > 10000:
        text = text[:10000]
    # append text to the corresponding tag
    tag_text_dict[tag] += text + ' '
    return tag



if __name__ == '__main__':
    source_dir = r'/Users/nird/Library/CloudStorage/OneDrive-Personal/Uni/שנה ב/מדעי הרוח הדיגיטליים/rashi/sources'
    output_file = f'outputs/words_stss_table_{make_timestamp()}.xlsx'

    # Iterate over files in source directory
    for root, _, files in os.walk(source_dir):
        for file in files:
            if file.endswith('.xlsx'):
                file_path = os.path.join(root, file)
                tag = extract_heb_text(file_path)
                print(f"Processed {file_path} with tag {tag}")


    # Save the text for each tag to a separate file
    for tag, text in tag_text_dict.items():
        if text:
            output_path = f'text_outputs/{tag}_text_{make_timestamp()}.txt'
            with open(output_path, 'w', encoding='utf-8') as f:
                f.write(text)
            print(f"Saved {tag} text to {output_path}")
        else:
            print(f"No text found for tag {tag}")
```

ד. מחברת פייתון ליצירת גרפים עבור טבלת השכיחויות:

https://1drv.ms/u/c/aa4c3e604c2a901d/EQzyyTf8gS1JhHhkWagY5M8BiZdROqAUCiKwVHpXhWmAJg?e=WpOmsK

נספח 4 – גרפים נוספים

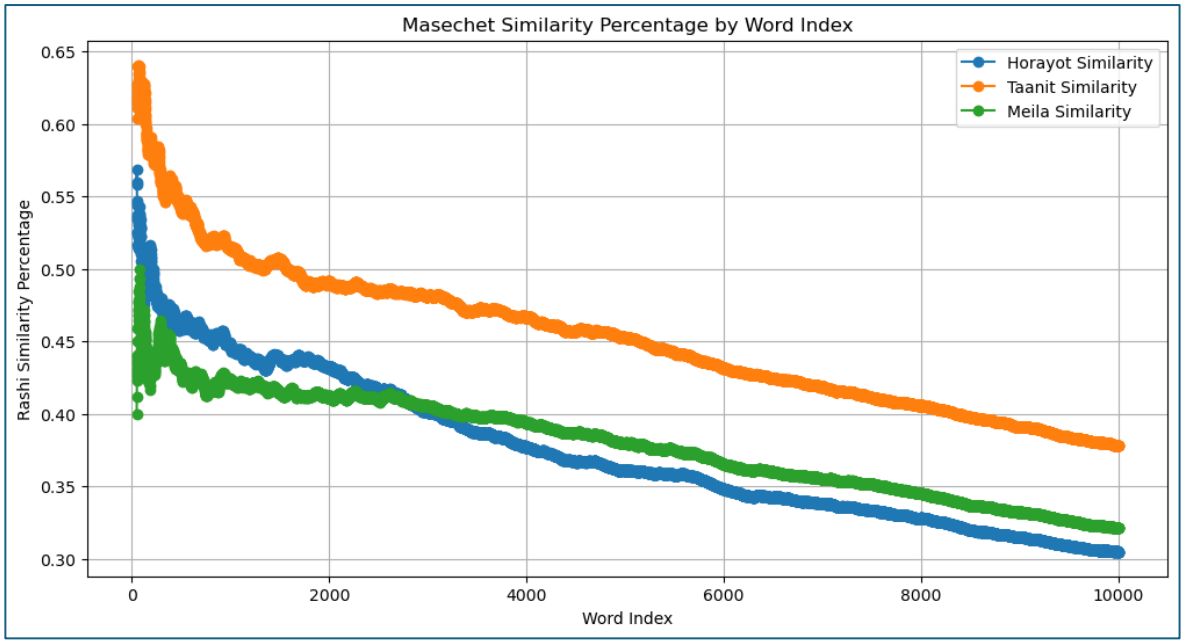
1. גרף heatmap התאמה לרש״י לפי מס׳ מילים השכיחות:



איור 8: גרף heatmap התאמה לרש״י לפי מס׳ מילים השכיחות.

הגרף צובע בירוק ערכים שקרובים לרש״י ובוורוד ערכים שרחוקים מרש״י. ניתן לראות כי מסכת תענית נשארת קרובה לאורך כל המשבצות, מסכת מעילה נשארת רחוקה לאורך כל המשבצות ומסכת הוריות ללא תוצאה חד-משמעית.

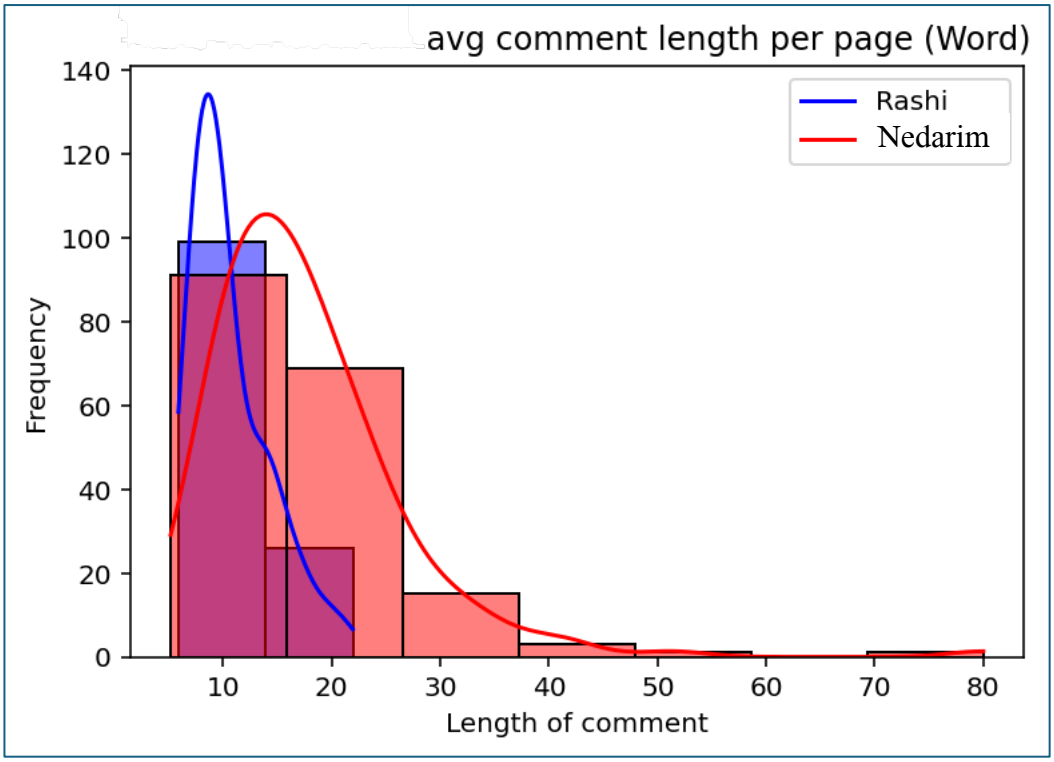
2. גרף פיזור המראה את מידת ההתאמה לרש״י עבור כל מסכת לפי מס׳ מילים שכיחות.



איור 9: גרף מידת ההתאמה לרש״י עבור כל מסכת לפי מס׳ מילים שכיחות.

הגרף מראה שמידת ההתאמה יורדת עבור כל המסכתות, בגלל שככל שמס׳ המילים רב יותר יש יותר מילים שהם נדירות וייחודיות לטקסטים השונים שמהם הרכבתי את האוסף של רש״י. לכן נכון יותר לבחור מרחב מדגם קטן בין 20-300 מילים.

3. התפלגות אורך הפירוש – רש"י מול פירוש מסכת נדרים:



איור 10: גרף התפלגות אורך הפירוש של פירוש רש"י מול פירוש מסכת נדרים.

הגרף מראה את צורת ההתפלגות של אורך הפירוש של רש"י לעומת הפירוש במסכת נדרים. ניתן לראות שרש"י בולט יותר בתחילת ציר ה-x (פרושים קצרים) והפירוש לנדרים נמשך יותר לצד ימין של הציר (פירושים ארוכים).

נספח 5 – טבלת השכיחויות המלאה

הטבלה מטה מראה את כל המילים שנכנסו לטבלת שכיחויות ואת השכיחויות שלהן בכל אוסף טקסטים.
מפאת גודל הקובץ, לא ניתן לצרף הטבלה למסמך. ניתן לצפות במסמך בקישור מטה:

https://1drv.ms/x/c/aa4c3e604c2a901d/EQBN-nJ9A3pKmB3pMas-K5ABMV2VlGTpe-eyV6_S7ZdK_w