בחינת סוגי ניתוחים תחביריים לתורה - דו"ח מחקר

תיאור שאלת המחקר

כיצד משתנים מאפייני התחביר והשפה בין ספרי התורה השונים ובין המקורות השונים לפי השערת התעודות, וכיצד ניתן לזהות מגמות עקביות ביניהם על סמך ניתוחים כמותיים.

המחקר מבקש לבחון האם קיימים הבדלים מובהקים בין ספרי התורה לבין המקורות השונים מבחינה תחבירית ולשונית, על סמך מדדים סטטיסטיים שונים.

סקירת ספרות

שיטות הניתוח התחבירי

ישנן שלוש שיטות מרכזיות לניתוח תחבירי בשפה העברית, בעיקר בלימודי התנ"ך והטקסטים המסורתיים:

- שיטת הפסוקיות – ניתוח על פי זיהוי פסוקיות במשפט.

שיטה זו מאפשרת להבין את הקשרים והדיוק התחבירי בין רכיבי המשפט בצורה ברורה משום שהיא מציגה את התהליך ההירכי ליצירת המשפט. באופן זה ניתן לזהות את הקשרים הלוגיים והסמנטיים בין החלקים. שיטה זו יעילה במיוחד במקרים של משפטים מורכבים בהם ישנם מספר חלקים תלויים זה בזה.

אחד החסרונות הגדולים של שיטה זו הוא שלעיתים ישנן בעיות במבני תחביר לא פורמליים או חופשיים, כמו בשפות טבעיות עם נטיות דיבור או שגיאות תחביריות. יש לכך השפעה על היכולת של המודלים לפענח את מבנה המשפט בצורה נכונה. בנוסף, במשפטים פשוטים יותר שיטה זו עלולה להיראות מורכבת מידי. המודל דורש ידע מעמיק במבנה הדקדוקי, והוא פחות אינטואיטיבי עבור קוראים שאינם בקיאים בעקרונות דקדוקיים פורמליים.

- שיטת התלויות – ניתוח על פי תלות מילים זו בזו.

שיטה זו מספקת דרך אינטואיטיבית ופשוטה לתאר יחסי תלות ישירים בין מילים במשפט. היתרון המרכזי שלה הוא ביכולת להדגיש את הקשרים התחביריים באופן ברור וחד-משמעי, מה שמועיל במיוחד בניתוח משפטים בעלי מבנים מורכבים. השיטה מתמקדת בקשרים הישירים בין המילים, ולכן שימושית בזיהוי יחסים בין ישויות, סיווג משפטים והבנת המשמעות.

חסרונה העיקרי הוא חוסר היכולת להתמודד עם מורכבות תחבירית גדולה או עם מבנים תחביריים מורכבים שלא ניתנים לתיאור בקלות באמצעות קשרי תלות ישירים בין מילים.

- שיטת טעמי המקרא – ניתוח על פי טעמי המקרא המשמשים גם כסימני פיסוק והכוונה מבנית.

טעמי המקרא הם שיטה מסורתית המשלבת בין קריאה, משמעות ותחביר. יתרונה המרכזי הוא בכך שהיא מספקת מבנה היררכי פשוט וברור המבוסס על חלוקות בינאריות, המשקף את הדגש והכוונה התחבירית של הפסוק כפי שהובנה במסורת היהודית.

עם זאת, היא מוגבלת במידת הדיוק התחבירי שלה, ואינה מתארת את כל היחסים התחביריים המורכבים בין המילים, במיוחד במבנים מודרניים יותר של תחביר.כמו כן, נדרשת הבנה טובה של סימני הטעמים ומשמעותם, מה שמגביל את הנגישות של השיטה.

השערת התעודות

השערת התעודות היא התאוריה שלפיה חמשת חומשי תורה נוצרו על ידי צירוף של מספר תעודות, שכל אחת מהן הייתה נרטיב עצמאי ושלם העומד בפני עצמו. התעודות הללו שיש ביניהן הקבלה מסוימת, נבדלות, לפי ההשערה, בסגנונן ובתוכניהן.

ההשערה פותחה במאות ה-18 וה-19, בניסיון להבין את הסתירות בתורה. לרוב מקובל לזהות ארבע תעודות, אך לב ההשערה איננו המספר המדויק של התעודות, אלא ההנחה שכל אחת מהן הייתה מסמך שלם. ההשערה שייכת לענף "הביקורת הגבוהה", חקר התפתחות ומקור הטקסטים, בתחום ביקורת המקרא.

השערת התעודות גורסת שהתורה מורכבת מטקסטים של 4 מקורות מרכזיים (J, E, D, P) בתוספת טקסט כלשהו שנוסף על ידי עורך אחד או יותר כדי להתאים את הטקסטים זה לזה (R):

- <u>המקור היהוויסטי (Jehovist – J)</u>: נכתב ב-950 לפנה"ס לערך בממלכת יהודה.

מאופיין בנטייתו לקצר בתיאורים והסברים, להצגת הקשר בין האל ובין האדם כישיר וכמעט מוחשי, ובשימוש בשם "יהוה" לכינוי האלוהים כבר מבריאת העולם (בניגוד למקורות E וP, המתאפיינים בשימוש נרחב בשם זה רק לאחר התגלות ה' למשה).

- <u>המקור האלוהיסטי (Elohist – E)</u>: נכתב ב-850 לפנה"ס לערך בממלכת ישראל.

נוטה לסיפורים עם מסר מוסרי, מתמקד בממלכת ישראל הצפונית.

- <u>המקור הדברימי (Deuteronomist – D)</u>: נכתב ב-600 לפנה"ס לערך, בתקופת הרפורמה הדתית של יאשיהו בירושלים.

מעריכים כי מקור זה מורכב מכמה שכבות. לפי סברה זו, בשלב הראשון (Dtn) כלל המקור את קובץ החוקים לבדו (ספר דברים, פרק י"ב-כו); בשלב השני (Dtr1), המתוארך לתקופת הרפורמה של יאשיהו, נערך הקובץ מחדש ונוספו לו ההקדמה הקושרת אותו לנאומי הפרידה של משה (ספר דברים, פרק א'-יא), ורשימת הברכות והקללות המובאת בסופו (ספר דברים, פרק כ"ז-ל). בתקופת גלות בבל נערך החיבור שוב (Dtr2), ונכללו בו קטעים נוספים, חלקם מבוססים על מקורות קדומים (למשל, שירת האזינו) וחלקם תוספות מאוחרות.

- המקור הכהני (Priestly – P): נכתב ב-500 לפנה"ס לערך, על ידי כהנים בגלות בבל.

המקור המפותח והארוך ביותר, והוא כולל את רוב ספר ויקרא ופרקים רבים בבראשית, שמות, ובמדבר. הוא מגלה עניין מיוחד בעבודת הקרבנות, בענייני טומאה וטהרה ובדקדוקי המצוות, ומתרחק מן התיאורים האנושיים של האל.

- המקור העורך (Redactor – R): מתייחס לעורך (או עורכים) שאסף, שילב וערך את ארבעת המקורות העיקריים של התורה (J, E, D, P) למסמך אחד, שהוא הטקסט שאנו מכירים כיום כתורה.
- מקור נוסף (Other – O): תוספת מודרנית שמטרתה להצביע על חלקים בעייתיים או יוצאי דופן מבחינת שיוך למקור ספציפי.

במיפוי הפסוקים של ספרי התורה לפי השערת המקורות, כל ספר משויך למספר מקורות בהתאם למאפייניו:

בראשית ושמות ממופים לפי ארבעת המקורות המרכזיים: J, E, P ו-R.

ויקרא מיוחס כמעט בלעדית למקור P, עם השפעות של R, מה שמעיד על אופיו החוקתי-פולחני.

במדבר ממשיך את הדפוס של בראשית ושמות, עם שילוב של J, E, P ו-R.

דברים כולל את המקור D בפיצוליו השונים (Dtr1, Dtr2, Dtn), לצד קטעים ממקור E, P, ו-O, מה שמעיד על עריכה מאוחרת יותר ושכבות טקסט שונות.

תרגול טעמים על טקסט מודרני

מחסור בכמה, ממוצרי החלב, של תנובה, החל, מהשבוע הבא, ובראשם, קוטג' תנובה.
המחלבה הודיעה, על, מחסור הצפוי, בחלק מהמוצרים, המיוצרים, במחלבת אלון-תבור, בשל, הפסקת ייצור, למספר ימים, בעקבות שדרוג, והחלפת, מערכות מחשוב.

עיקר המחסור, צפוי, בגבינת קוטג', עקב, חיי מדף, קצרים.
לפי ההערכות, יהיו מעט, גביעי קוטג', במדפים, בשבוע הבא.
בתנובה, לא מסרו, הערכה, עד מתי, יימשך המחסור.

למרות, שהפסקת הייצור, תהיה, במוצרים אחרים, המחלבה נערכה, עם מלאי, ואף, הגבירה ייצור, של, מוצרי חלב, שמחיריהם, נמצאים בפיקוח.

לא צפוי, מחסור, בשמנת חמוצה, ולבן, ולא, בגבינת נפוליאון.
לא צפוי, מחסור, בשמנת מתוקה, שסבלה, ממחסור ממושך, לאחרונה, משום, שהיא מיוצרת, במחלבת רחובות.
מחסור צפוי, ביוגורטי יופלה, שאינם גו, בעוד כשבועיים.

בחינת השפעת הניתוח התחבירי על פרשנות הפסוק

נראה 10 פסוקים שמשמעותם תלויה בחלוקה התחבירית שלהם, ונבחן את משמעות הפסוק על פי שלושת סוגי הניתוח התחבירי.

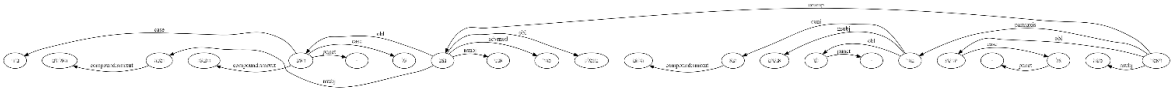
עבור הפסוק הראשון נציג את הפירוט המלא לכל ניתוח תחבירי, לאחר מכן עבור שאר הפסוקים נציג את המסקנה המתקבלת.

שמות יז, ט: "וַיֹּאמֶר מֹשֶׁה אֶל-יְהוֹשֻׁעַ בְּחַר-לָנוּ אֲנָשִׁים וְצֵא הִלָּחֵם בַּעֲמָלֵק מָחָר אָנֹכִי נִצָּב עַל-רֹאשׁ הַגִּבְעָה וּמַטֵּה הָאֱלֹהִים בְּיָדִי"

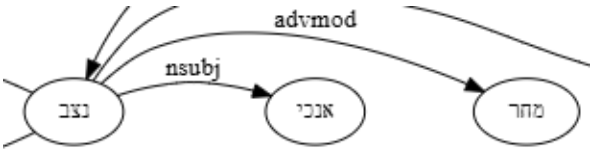
<u>מבנה פסוקיות:</u>
הבחינה של מבנה הפסוק עוסקת בחלוקה הפנימית של הפסוק ובאופן שבו המילים מסודרות בצורה תחבירית. במובן זה, לא תמיד המשמעות משתנה באופן דרמטי, אך יכולת ההבנה של כוונת המחבר או של סדר הפעולות משתפרת. לדוגמה נבחין, שתחת הפסוקית עם המזהה n434479 נמצאות המילים מחר ונצב. כאשר אנו מזהים את הציווי של משה (ליהושע) ואת הפעולה של משה עצמו, מבנה הפסוק יכול להבהיר את הדיאלוג בין הדמויות – משה נותן הוראה ליהושע, בעוד שמשה עצמו ממקם את עצמו באופן עתידי (מחר) בראש הגבעה.

<u>מבנה תלויות:</u>

בהרצת המודל של דיקטה נקבל:



ובתצוגה ממוקדת על המילים הרלוונטיות:



ההבנה של מבנה התלויות יכולה לשפוך אור על הדרך שבה כל מרכיב משפיע על שאר המרכיבים של הפסוק. כאשר רואים שהפעולה של משה תלויה בסמכות האלוהית שמופיעה בידו (ה"נִצָּב" על הגבעה עם "מַטֵּה הָאֱלֹהִים בְּיָדִי"), ניתן להבין את חוויית הקיום השונה של כל דמות. השימוש במילה "מָחָר" יוצר תכנון של זמן. כלומר נצב הוא הפועל הראשי שקשור למחר, כך שמשה מדבר על עמידתו העתידית על הגבעה.

טעמי המקרא:

המילה מחר נמצאת תחת הקיסרות השינה, וכך משפיעה על המילה ניצב. כלומר גם כאן משה מדבר על עמידתו העתידית על הגבעה.
הבחינה של עצי הטעמים עשויה לשנות את המשמעות המילולית על ידי הצגת הקשרים המילוליים בין חלקי הפסוק, וזה יכול להוביל להבנה טובה יותר של הכוונה הרוחנית של הפסוק.
למשל, השימוש ב"אָנֹכִי נִצָּב" מצביע על נוכחות פעילה של משה, אך הקשר בינו לבין "מַטֵּה הָאֱלֹהִים" מדגיש את ההיבט הרוחני של הקרב – קרב שבו כוח אלוהי ויכולת מנהיגותית של משה משתלבים.
מעבר לכך, הקשר בין המילים מחדד את השפעתם של כל המרכיבים (הקרב, הזמן, הסמכות) על כלל התמונה.

בראשית לד, ז: "וּבְנֵי יַעֲקֹב בָּאוּ מִן-הַשָּׂדֶה כְּשָׁמְעָם וַיִּתְעַצְּבוּ הָאֲנָשִׁים וַיִּחַר לָהֶם מְאֹד כִּי-נְבָלָה עָשָׂה בְיִשְׂרָאֵל לִשְׁכַּב אֶת-בַּת-יַעֲקֹב וְכֵן לֹא יֵעָשֶׂה"

מבנה פסוקיות:

- "וּבְנֵי יַעֲקֹב בָּאוּ מִן-הַשָּׂדֶה" – פתיחה שמציינת את המיקום של בני יעקב ואת הפעולה שהם עושים, שהם באים מהשדה.
- "כְּשָׁמְעָם" – מילה שמצביעה על הזמן שבו שמעו את החדשות, כלומר, מיד כשהם שמעו.
- "וַיִּתְעַצְּבוּ הָאֲנָשִׁים" – פועל המתאר את תחושת הצער של בני יעקב בעקבות החדשות.
- "וַיִּחַר לָהֶם מְאֹד" – הוספת רגש של כעס וחומרת האירוע.
- "כִּי-נְבָלָה עָשָׂה בְיִשְׂרָאֵל" – הסיבה לתגובה של בני יעקב; הם מתעצבנים כי נעשה דבר נורא, נבלה.
- "לִשְׁכַּב אֶת-בַּת-יַעֲקֹב" – פירוט מה קרה, תיאור של המעשה הנורא.
- "וְכֵן לֹא יֵעָשֶׂה" – הצהרה מוסרית שמבהירה שדבר כזה לא יכול לקרות, שזה לא מקובל.

כלומר מבנה פסוקיות עוזר להבין את סדר הפעולות והרגשות של בני יעקב ואת הקשר בין שמיעת החדשות לתגובה הרגשית.

מבנה תלויות:

- "וּבְנֵי יַעֲקֹב בָּאוּ מִן-הַשָּׂדֶה" – זהו תיאור של הפעולה הפיזית של בני יעקב, המובילה לתגובה.
- "כְּשָׁמְעָם" – יש תלות בין שמיעת החדשות ובין הרגש שהולך ומתעורר אצל בני יעקב.
- "וַיִּתְעַצְּבוּ וַיִּחַר לָהֶם מְאֹד" – התגובה הרגשית היא ישירה לשמיעה, כלומר, הרגש תלוי ישירות במידע שנמסר להם.

- "כִּי-נְבָלָה עָשָׂה בְיִשְׂרָאֵל" – הסיבה לכעס ולצער של בני יעקב היא המעשה שנעשה, וההבנה שדבר כזה אסור.
- "לִשְׁכַּב אֶת-בַּת-יַעֲקֹב" – סיבה ישירה למה שהם זעמו – פגיעה בכבוד המשפחה.
- "וְכֵן לֹא יֵעָשֶׂה" – מסקנה מוסרית מתבקשת מהאירועים – כזה מעשה לא ייעשה.

כלומר מבנה תלויות מדגיש את הקשרים בין כל אחד מהמרכיבים: שמיעת החדשות, התחושות, העשייה והמוסר.

טעמי המקרא:

- "וּבְנֵי יַעֲקֹב בָּאוּ מִן-הַשָּׂדֶה" – ההתחלה של הפסוק מציינת את המיקום (השדה) ותנאים של הזמן (החזרה מהשדה), כך שהמילים "בָּאוּ" ו"מִן-הַשָּׂדֶה" מקשרות בין הפעולה הגשמית לבין שאר התגובה.
- "כְּשָׁמְעָם" – צורת המילים משדרת את המיידיות של הפעולה. הם שמעו את החדשות, והתגובה לא מאחרת לבוא.
- "וַיִּתְעַצְּבוּ" ו-"וַיִּחַר לָהֶם מְאֹד" – הקשר בין שני הפעלים מדגיש את עוצמת התגובה: עצב מאוד חזק שמתערבב בכעס.
- "כִּי-נְבָלָה עָשָׂה בְיִשְׂרָאֵל" – המילה "נְבָלָה" מתארת את המעשה הנורא בצורה מאוד ברורה, והצמדת המילה "עָשָׂה" מדגישה את פעולתו של האיש שעשה את המעשה.
- "לִשְׁכַּב אֶת-בַּת-יַעֲקֹב" – המילים "לִשְׁכַּב" ו"אֶת-בַּת-יַעֲקֹב" מבארות את הבעיה הספציפית – חטא של חיבור לא ראוי.
- "וְכֵן לֹא יֵעָשֶׂה" – סיום הצהרה מוסרית המחברת בין המעשה לבין העיקרון: דבר כזה לא ייעשה.

כלומר עצי טעמים מקנים את הממד המוסרי והרגשי של הפסוק, תוך הדגשת המעשה הנורא ותחושת הכעס והצער שמורגשות בעקבותיו.כלומר, כל מילה מקיימת קשרים עם אחרות שמחברות את התגובה המוסרית והרגשית.

מסקנות לניתוח שאר הפסוקים:

לאחר שניתחו את שני הפסוקים הנ"ל, הבחנו כי אין שוני במשמעות הכוללת מבחינת שלושת סוגי הניתוח התחביריים, אך כל אחד מהניתוחים חושף היבט אחר של הפסוק:

- מבנה פסוקיות נותן תמונה תחבירית ברורה של התקדמות הפעולה בין הדמויות.
- מבנה תלויות מדגיש את הקשרים הדינמיים בין כל אלמנט, כך שניתן להבין את הקרב כשלב בשרשרת פעולות שכוללות זמן, ציווי וסמכות.
- עצי טעמים מציגים את המשמעות העמוקה יותר של המילים, במיוחד בהקשרים רוחניים, כמו השפעת כוח אלוהי במלחמה.

כלומר כל אחת מהגישות שואפת להדגיש פרספקטיבה שונה שמעשירה את ההבנה הכוללת של הפסוק. לכן עבור הפסוקים הנותרים, בחרנו להתמקד בהיבטים המשמעותיים השונים המתקבלים על ידי הניתוחים התחביריים.

עבור כל אחד מהפסוקים ביצענו ניתוח באופן זהה למפורט למעלה, לצורך מניעת חזרתיות בחרנו להוסיף לדוח רק את המסקנות שהתקבלו מניתוח זה.

שמות כ, ב: "אָנֹכִי יְהוָה אֱלֹהֶיךָ אֲשֶׁר הוֹצֵאתִיךָ מֵאֶרֶץ מִצְרַיִם מִבֵּית עֲבָדִים לֹא-יִהְיֶה לְךָ אֱלֹהִים אֲחֵרִים עַל-פָּנָי"

- מבנה פסוקיות: מבנה פסוקי כאן מבדל בין שני חלקים עיקריים: הצהרת הוויה של אלוהים וההנחה שהוא הוציא את בני ישראל ממצרים, והציווי המוסרי לא לקבל אלים אחרים.
- מבנה תלויות: יש קשר בין פעולתו של אלוהים לבין דרישת הנאמנות של העם. זו תלות בין החירות לבין הציווי הדתי.
- עצי טעמים: כל מילה ותגובה בפסוק מדגישות את חשיבות הנאמנות לאל אחד בלבד. כל החלקים מתחברים יחד ומצביעים על סמכותו של אלוהים הייחודי ודרישתו מהעם לא לעבוד אלים אחרים.

בראשית א, טז: "וַיַּעַשׂ אֱלֹהִים אֶת-שְׁנֵי הַמְּאֹרֹת הַגְּדֹלִים אֶת-הַמָּאוֹר הַגָּדֹל לְמֶמְשֶׁלֶת הַיּוֹם וְאֶת-הַמָּאוֹר הַקָּטֹן לְמֶמְשֶׁלֶת הַלַּיְלָה וְאֵת הַכּוֹכָבִים"

- מבנה פסוקיות: הוא מציג את פעולתו של אלוהים בצורה ברורה ומסודרת: יצירת שני המאורות הגדולים לשם שליטה ביום ובלילה.
- מבנה תלויות: יש קשר ברור בין כל אחד מהמאורות לבין הזמן: השמש ליום והירח ללילה.
- עצי טעמים: כל מילה בפסוק מציינת את תפקידו של כל מאור ומייחסת עבור כל אחד מהם שליטה על תקופת זמן מסוימת.

בראשית טו, יג: "וַיֹּאמֶר לְאַבְרָם יָדֹעַ תֵּדַע כִּי-גֵר יִהְיֶה זַרְעֲךָ בְּאֶרֶץ לֹא לָהֶם וַעֲבָדוּם וְעִנּוּ אֹתָם אַרְבַּע מֵאוֹת שָׁנָה"

- מבנה פסוקותי: מבנה תחבירי ברור, שבו ההצהרה על ידיעת אברהם באה לפני תיאור העתיד של בניו.
- מבנה תלויות: יש קשר ברור בין הידיעה של אברהם לבין העתיד הצפוי לבניו: עבדות, עינוי, ושעבוד במצרים.
- עצי טעמים: כל מילה בונה את התמונה המורכבת של הסבל העתידי: הזרות, השעבוד, העינוי, והזמן המוגדר.

דברים ו, ז: "וְשִׁנַּנְתָּם לְבָנֶיךָ וְדִבַּרְתָּ בָּם בְּשִׁבְתְּךָ בְּבֵיתֶךָ וּבְלֶכְתְּךָ בַדֶּרֶךְ וּבְשָׁכְבְּךָ וּבְקוּמֶךָ"

- מבנה פסוקיות: מציין סדר של פעולות חינוכיות הנעשות בזמנים ובמיקומים שונים.
- מבנה תלויות: כל פעולה תלויה בזמן ובמקום: הבית, הדרך, השכיבה והקימה.

- עצי טעמים: המילים והפעלים מחזקים את הצורך בהדרכה ושינון התורה בצורה רציפה ואחידה, בכל זמן ובכל מקום.

ויקרא כג, טז: "עַד מִמָּחֳרַת הַשַּׁבָּת הַשְּׁבִיעִת תִּסְפְּרוּ חֲמִשִּׁים יוֹם וְהִקְרַבְתֶּם מִנְחָה חֲדָשָׁה לַיהוָה"

- מבנה פסוקיות: תחילת הזמן (ספירת העומר) עד למועד הקרבת המנחה, כל פעולה מתקיימת בזמן מוגדר ועם סיבה מסוימת.
- מבנה תלויות: יש קשר הדוק בין פעולת הספירה לבין פעולת הקרבת המנחה החדשה: השניים מבוצעים כחלק מההכנה לחג השבועות.
- עצי טעמים: דגש על הציווי הממוקד והמדוד לספור את הימים, והקרבה של מנחה חדשה באירוע מרכזי זה.

בראשית יג, יג: "וְאַנְשֵׁי סְדֹם רָעִים וְחַטָּאִים לַיהוָה מְאֹד"

- מבנה פסוקיות: תחילת הפסוק מציינת את זהותם של אנשי סדום, וממשיך לתאר את אופיים הרע והחטא שלהם כלפי אלוהים.
- מבנה תלויות: יש קשר בין הרוע המוסרי לבין החטא הדתי, והפסוק מבהיר שהחטא לא היה רק פוגע בחברה אלא גם בריבונות של אלוהים.
- עצי טעמים: החיבור בין "רָעִים" ו"חַטָּאִים" ו"לַיהוָה מְאֹד" מחדד את החומרה המוסרית והדתית של המעשה.

שמות א, א: "וְאֵלֶּה שְׁמוֹת בְּנֵי יִשְׂרָאֵל הַבָּאִים מִצְרָיְמָה אֵת יַעֲקֹב אִישׁ וּבֵיתוֹ בָּאוּ"

- מבנה פסוקיות: הפסוק מציין את שמות בני ישראל ומקשר את הגעת יעקב ובני ביתו למצרים.
- מבנה תלויות: יש קשר ישיר בין השמות של בני ישראל לבין הגעת משפחת יעקב.
- עצי טעמים: כל מילה מבהירה את מקומם של בני ישראל ויוצרים הקשר ברור של משפחת יעקב בתור ציר חשוב.

ויקרא כא, א: "וַיֹּאמֶר יְהוָה אֶל-מֹשֶׁה אֱמֹר אֶל-הַכֹּהֲנִים בְּנֵי אַהֲרֹן וְאָמַרְתָּ אֲלֵהֶם לְנֶפֶשׁ לֹא-יִטַּמָּא בְּעַמָּיו"

- מבנה פסוקיות: הפסוק בנוי משני חלקים: הציווי האלוהי למשה, וההוראות למשה להעביר את הציווי לכוהנים.
- מבנה תלויות: יש קשר ברור בין הדיבור האלוהי, הפעולה של משה בהעברת הציווי, והחובה של הכוהנים לשמור על טהרתם.
- עצי טעמים: כל מילה מדגישה את הציווי המוחלט, את חשיבות הקדושה וההפרדה של הכוהנים מהעם על מנת לשמור על קדושתם.

נתונים סטטיסטיים כלליים וניתוחם

המילים ושכיחותן

10 המילים השכיחות ביותר הן:

אֶת	2299
אֶל	1091
עַל	856
כָּל	790
וְאֶת	607
אֲשֶׁר	445
אֲשֶׁר	366
יְהוָה	317
כִּי	312
אֲשֶׁר	284

שימוש תדיר במילות יחס וחיבור

רוב המילים המוזכרות הן מילים תפקודיות, כלומר מילים שאין להן משמעות עצמאית אך הן חיוניות להבהרת קשרים תחביריים ולקישור בין חלקי המשפט.

שכיחותן הגבוהה מצביעה על כך שהשפה המקראית משתמשת במבנים תחביריים מורכבים.

החזרה הרבה של מילים כמו את (2299 פעמים) ו-אֶת (607 פעמים) מצביעה על החשיבות הרבה שלהן בתקשורת ובחיבור בין חלקי הטקסט. זה לא מפתיע כי מילים כאלה נדרשות לעיתים קרובות בכתיבה תחבירית.

הדגשה על הכללה ופרטים

המילה כָּל (790 פעמים) מופיעה בהקשרים של הכללה, דבר המעיד על הנטייה של הטקסט להדגיש קבוצה כוללת או מצבים גורפים (למשל, "כָּל־הָעָם", "כָּל־הַבְּהֵמָה").

השכיחות הגבוהה של מילה זו עשויה גם להצביע על עיסוק חוזר בחוקים, דינים ותיאורים כוללים של עם ישראל והחוקים החלים על כולם.

מילות חיבור ותנאי

אֲשֶׁר (445, 366 ו-284 פעמים) היא מילת זיקה שחוזרת בשלוש וריאציות שונות של ניקוד. ההבדלים בתווים הדיאקריטיים (ניקוד וטעמים) יוצרים מראית עין של מילים שונות, אך למעשה מדובר באותה מילה. השכיחות שלה משקפת את הסגנון המורכב של הטקסט, המבוסס על משפטים ארוכים ומורכבים.

ההבדלים בין אֲשֶׁר ל-אֲשֶׁר יכולים להיות תוצאה של גירסאות דקדוקיות שונות (לדוגמה, צורת הכתיבה וההגייה המקובלת בכל תקופה או אזור).

כִּי (312 פעמים) משמשת כמילת סיבה או תנאי. שכיחותה הגבוהה מצביעה על כך שהרבה משפטים במקרא מתארים סיבות, תנאים או תוצאות של מעשים.

היבטים תיאולוגיים

יְהוָה (317 פעמים): זהו שם הוויה של אלוהים, והוא מופיע רבות בטקסטים הדתיים. שכיחות גבוהה זו משקפת את מרכזיותו של אלוהים בסיפורים המקראיים ובחוקים.

מילה זאת מופיעה במגוון סוגים נוספים של וריאציות בתנ"ך דבר המחזק את שכיחותה הגבוהה:

יְהוָה: 317, יְהוָהֹ: 166, יְהֹוָה: 158, יְהוֶה: 146, יְהוִה: 143, יְהוֹוָה: 124, יְהוֶה: 99, יְהוֹוָה: 85, יְהֹוָה: 69, יְהוָה: 56, יְהֹוָה: 38, יְהוָה: 36, יְהֹוָה: 34, יְהוָהֹ: 16, יְהֹוָה: 8, יְהוֹוָה: 6, יְהוָהֹ: 3, יְהֹוָה: 3, יְהוָהֹ: 3, יְהוֹוִה: 2, יְהֹוָה: 2, יְהֹוָה: 1.

נתונים סטטיסטיים וניתוחם בפילוח לפי ספר

אורכי פסוקים

ממוצע אורך פסוק באופן כללי: 13.66 מילים

ממוצע אורך פסוק לפי ספר:

דברים	14.94
ויקרא	13.91
שמות	13.81
בראשית	13.42
במדבר	12.70

באופן כללי, כל הספרים נעים בטווח של 12.7-14.94 מילים לפסוק, כך שיש עקביות יחסית באורך הפסוקים (לא מדובר בשינויים קיצוניים). זה מעיד על סגנון כתיבה אחיד יחסית לאורך הספרים, במיוחד כאשר מדובר בספרים דתיים כמו התורה.

קיימת נטייה לפסוקים ארוכים יותר בספר דברים וקצרים יותר בספר במדבר.

מספר פסוקיות בפסוק

ממוצע של פסוקיות בפסוק לפי ספר:

על פי פסוקיות הטעמים		על פי מבנה הפסוקיות	
דברים	1.948745	דברים	4.245055
שמות	1.940496	בראשית	3.978738
ויקרא	1.939464	שמות	3.782343
בראשית	1.935463	ויקרא	3.733010
במדבר	1.890528	במדבר	3.358733

<u>הבדלים במבנה הפסוקים בין שני סוגי הניתוחים:</u>

<u>פסוקיות הטעמים:</u>

- ממוצע מספר הפסוקיות בפסוק נע בין 1.89 ל-1.95, דבר המעיד על מבנה תחבירי פשוט יחסית.
- הטעמים מייצגים את הפיסוק המסורתי של הטקסט המקראי, ולכן נוטים לפצל את הפסוקים לפחות יחידות תחביריות.
- ספר דברים מציג את הממוצע הגבוה ביותר ובמדבר את הממוצע הנמוך ביותר.

<u>מבנה הפסוקיות:</u>

- הממוצעים גבוהים בהרבה (3.35–4.25), מה שמעיד על רמת פירוט תחבירי גבוהה יותר בניתוח זה.
- כאן, הפסוקים מפורקים למבנים תחביריים ברורים כמו נושא, נשוא ומשלים, דבר שמסביר את עליית מספר הפסוקיות.
- גם כאן, ספר דברים מוביל ובמדבר הנמוך ביותר.

מסקנות

הפערים בין שתי השיטות מצביעים על הבדל בין פיסוק מסורתי (טעמים) לבין ניתוח תחבירי מפורט (מבנה פסוקיות).

ספר דברים מציג את הממוצע הגבוה ביותר, מה שמעיד על מורכבות התחביר שבו, המתבטאת גם בפיסוק וגם במבנה התחבירי.

ספר במדבר נוטה להכיל פסוקים קצרים ופשוטים יותר מבחינה תחבירית

ניתן לראות שאין עקביות בספרים שמות, ויקרא ובראשית מבחינת סדר הנתונים הממוין בין שתי השיטות.

עומקי פסוקים

ממוצע של עומקי פסוקים לפי ספר:

על פי עץ הטעמים		על פי עץ תלויות	
דברים	2.156904	דברים	6.280501
ויקרא	2.135041	שמות	6.064303
בראשית	2.132334	ויקרא	6.041909
שמות	2.131405	בראשית	6.005871
במדבר	2.059783	במדבר	5.992242

השוואה בין עץ הטעמים לעץ התלויות:

ממוצע עומקי הפסוקים לפי עץ הטעמים ממוצע נע בין 2.05 ל-2.16.

ממוצע עומקי הפסוקים לפי עץ התלויות נע בין 5.99 ל-6.28, ערכים גבוהים משמעותית מעומקי הטעמים.

עץ התלויות חושף מורכבות תחבירית עמוקה יותר לעומת עץ הטעמים, כפי שניתן לראות מהפער הגדול בין הממוצעים.

דפוסים בין הספרים:

ספר דברים מציג את העומק הממוצע הגבוה ביותר בשתי השיטות, וספר במדבר מציג את העומק הממוצע הנמוך ביותר בשתי השיטות, בשאר הספרים (בראשית, שמות, ויקרא), העומק דומה יחסית בשתי השיטות ונע בטווחים קרובים (2.13-2.14 בעץ הטעמים ו-6.00-6.06 בעץ התלויות). אין עקביות מבחינת סדר הנתונים הממוין בין שתי השיטות.

מסקנות

כל הספרים מציגים ממוצע עומק פסוקים דומה מאוד בעץ טעמים (~2) ובעץ תלויות (~6) , ללא הבדל משמעותי ביניהם.

ממוצע עומק הפסוקים אינו כלי מובהק להפרדת הספרים, כיוון שהוא זהה בכל המקורות שנבדקו.

הפסוקיות ושכיחותן

תדירות הפונקציות הדקדוקיות בספרים השונים

ספר בראשית מוביל בשימוש בנשוא (Pred), נושאים (Subj), ושאלות (Ques), מה שמעיד על נרטיב עשיר ומורכב יותר.

ספר ויקרא מכיל פחות פונקציות דקדוקיות כמו שאלות (Ques) ותחביר רגשי (Intj, Exst), מה שעשוי להצביע על סגנון הצהרתי ורשמי יותר.

ספרי דברים ושמות מכילים שימוש נרחב במבני חיבור (Conj), מה שמעיד על מבנים תחביריים מורכבים יותר או חזרתיות גבוהה של מבנים מקשרים.

ספר במדבר דומה במבנהו לשמות, עם שימוש גבוה יחסית בנשואים ובמבני קשר.

תדירות סוגי הביטויים

בראשית ושמות הם הספרים עם הכי הרבה ביטויי שם עצם (NP) וביטויי פועל (VP), מה שמעיד על נרטיב עשיר ודגש על פעולות ודמויות. בנוסף, הספרים מכילים גם הרבה ביטויי צירוף (CP), דבר שיכול להעיד על משפטים מורכבים יותר.

ויקרא מראה שימוש מופחת ב-VP (ביטויי פועל) בהשוואה לשאר הספרים, מה שמרמז על סגנון חקיקתי ופחות נרטיבי.

בספר במדבר יש שימוש ניכר בביטויי יחס (PP), מה שמעיד על רשימות ארוכות ותיאורי מסעות. הספר מכיל איזון בין נרטיב לרשימות, ולכן מגוון התחביר נמוך מעט משמות אך גבוה מויקרא.

בספר דברים ניכר שימוש גבוה בביטויי שם עצם (NP), מה שמעיד על מבנים תחביריים מחוזקים ואפקטיביים להעברת מסרים.

מסקנות

בראשית הוא הספר המורכב ביותר תחבירית, עם הכי הרבה ביטויים תחביריים (17,981), מה שמתאים לאופיו הסיפורי העשיר.

ויקרא הוא הספר עם הכי פחות ביטויים תחביריים (9,619), מה שמעיד על מבנה לשוני פשוט וחזרתי, כנראה בשל אופיו החקיקתי.

שמות ובמדבר מכילים איזון בין סגנון נרטיבי לחקיקתי, ולכן מספר הביטויים שלהם נמצא בין בראשית לויקרא.

דברים מציג סגנון ייחודי עם הרבה ביטויי חיבור, דבר שמתאים לאופיו הרטורי והשכנועי.

מילים ייחודיות לפי ספר

ספר	סך כל המילים בספר	מילים ייחודיות בספר	אחוזים
דברים	14280	7818	54.747899
בראשית	20573	10675	51.888397
שמות	16710	8446	50.544584
ויקרא	11950	5817	48.677824
במדבר	16374	7847	47.923537

ספר דברים הוא הספר עם אחוז המילים השונות הגבוה ביותר, מה שמעיד על סגנון רטורי עם פחות חזרות ישירות. כנראה שהדגש כאן הוא על נאומים והוראות כלליות, ולא על חוקים פורמליים חוזרים כמו בויקרא.

ספר בראשית הוא הספר עם אוצר המילים הגדול ביותר, מה שמעיד על נרטיב עשיר ומגוון לשונית. עם זאת, אחוז המילים השונות בינוני, כלומר יש חזרות מסוימות על ביטויים ודמויות מרכזיות.

בספר שמות אחוז המילים השונות נמוך מעט מבראשית, מה שמעיד על יותר חזרות על מבנים לשוניים קבועים. סביר להניח שהחזרות קשורות לתיאורי יציאת מצרים ולחוקים ראשוניים שנמסרו לעם.

ספר ויקרא זהו הספר עם אחוז המילים השונות הנמוך ביותר, מה שמעיד על חזרתיות גבוהה מאוד. מתאים לכך שוויקרא הוא בעיקר ספר חוקים וטקסים, החוזרים בניסוחיהם שוב ושוב.

בספר במדבר אחוז המילים השונות נמוך יחסית, כנראה בגלל ריבוי רשימות, חוקים וסיפורי מסעות חוזרים. מכיל שילוב בין נרטיבים לבין חלקים מחוקקים, ולכן מציג תמהיל בין חזרות למגוון מילים.

<u>מסקנות</u>

בממוצע, כ-50% מהמילים בספרים חוזרות על עצמן, מה שמעיד על שילוב של חזרות סגנוניות וגיוון מתון. ספר דברים מציג את האחוז הגדול ביותר, ספר במדבר מציג את האחוז הקטן ביותר.

לא מדובר בטקסט שכתוב בשפה חופשית כמו בספרות מודרנית – יש בו מבנים לשוניים מקובעים, מה שמוביל ליציבות סטטיסטית.

ביצענו גם ניתוח סטטיסטי על מילים ייחודיות לפי כל פסוק בספר, האחוזים שהתקבלו שאפו ל-100% בכל אחד מהספרים ולכן לא ניתן להסיק מסקנות רלוונטיות מהנתון הזה. ניתוח הנתונים לפי כל פסוק (משפט קצר יחסית) יכול לגרום לכל מילה להיחשב ייחודית.

אורך מילים בספרים

ממוצע של אורכי מילים לפי ספר:

במדבר	8.11
דברים	8.04
שמות	7.99
בראשית	7.91
ויקרא	7.81

אורך המילים דומה בין רוב הספרים: הממוצעים של אורך המילים לא שונים בצורה דרמטית בין הספרים, ונעים בטווח של 7.81 עד 8.11 אותיות למילה. הדבר מצביע על כך שמבנה המילים בתנ"ך באופן כללי די אחיד, לפחות בהיבט של אורך המילים. כלומר, אין שינוי משמעותי לאורך המילים בין ספרי התורה.

קיים קשר בין אורך המילים לאופי התוכן. ההבדלים בין במדבר וויקרא יכולים לשקף את האופי השונה של התוכן בספרים:

ספר במדבר מציג את המילים הארוכות ביותר.

ויקרא מציג את המילים הקצרות ביותר. הספר הזה מתמקד בעיקר בציוויים דתיים, אשר לרוב מבוטאים בשפה תמציתית וממוקדת יותר, כך שמילים קצרות יותר עשויות להיות חלק מהתכנים הקצרים והישירים.

תבניות הפסוקים

ביצענו ניתוח סטטיסטי לפי תבניות הפסוקיות אך לא הגענו למסקנות רלוונטיות. בנוסף, נותחו מבני העצים על פי המילים וההקשרים שלהם, ולכן הוספת ייצוג נוסף ללא מילים עלולה להעמיס מידע נוסף מבלי לספק תובנות חדשות, מה שעלול להפוך את הניתוח לפחות ממוקד ויעיל ולכן בחרנו לא להכניס ניתוח זה לסיכום.

מסקנות סופיות - נתונים סטטיסטיים בפילוח לפי ספר

ניתן לראות שבניתוח נתונים לפי אורכי פסוקים, מספר פסוקיות בפסוק בשתי הדרכים (מבנה פסוקיות ופסוקיות טעמים), עומקי פסוקים לפי שני הייצוגים (עץ טעמים ועץ תלויות) ומילים ייחודיות לפי ספר- בכולם ספר דברים מציג את הנתונים הגבוהים ביותר וספר במדבר מציג את הנתונים הנמוכים ביותר. בניגוד לכך, ניתן לראות שבניתוח אורך מילים לפי ספר במדבר נמצא במקום הגבוה ביותר.

עיקרו של ספר דברים הוא נאומו של משה בו הוא סוקר את ההיסטוריה של עם ישראל ואת החוקים שקיבל, ולכן המשפטים בו עשויים להיות ארוכים יותר כדי לתאר את האירועים והדינמיקה.

ואילו בספר במדבר מצויינים רשימות רבות וחוקים קצרים, הוא בעל הוראות דתיות וצבאיות שהם לרוב פשוטים תחבירית כנראה בגלל התיאורים הטכניים והרשימות שבו.

<u>נתונים סטטיסטיים וניתוחם בפילוח לפי מקור</u>

אורכי פסוקים

ממוצע אורך פסוק באופן כללי: 13.66 מילים
ממוצע אורך פסוק לפי מקור:

D2	17.711538
Dn	15.187683
D1	14.970402
E	14.347413
J	13.571025
P	13.270426
O	11.636364
R	11.153310

<u>מסקנות</u>

הבדלים בין מקורות: ניתן לראות ש-D2 בעל ממוצע אורך הפסוק הגבוה ביותר בעוד ש-R בעל ממוצע אורך הפסוק הנמוך ביותר.

פשטות/מורכבות של מקורות שונים: מקורות עם ממוצע נמוך יותר של אורך פסוק (כגון R, O) עשויים להיות קלים יותר להבנה או עם יותר משפטים פשוטים, בעוד שמקורות עם ממוצע גבוה יותר (כגון D2) עשויים להציג מבנים תחביריים מורכבים יותר או משפטים ארוכים ומפורטים.

מספר פסוקיות בפסוק

ממוצע של פסוקיות בפסוק לפי מקור:

על פי פסוקיות הטעמים		על פי מבנה הפסוקיות	
D2	1.250000	D1	1.771564
O	1.244300	D2	1.625000
E	1.231795	Dn	1.527473
J	1.231281	O	1.480769
Dn	1.227914	P	1.451100
D1	1.222759	R	1.395176
P	1.207273	J	1.299563
R	1.204647	E	1.272957

<u>הבדלים במבנה הפסוקים בין שני סוגי הניתוחים:</u>

<u>פסוקיות הטעמים:</u>

- כל המקורות מציגים ממוצע שנע בין 1.2 ל-1.25. אין פערים משמעותיים בין רוב המקורות בממוצע זה, כך שייתכן שמדובר על תבנית תחבירית דומה לאורך רוב הספרים בתנ"ך.
- המקור D2 בעל הממוצע הגבוה ביותר של פסוקיות לפי טעמים והמקור R בעל הממוצע הנמוך ביותר של פסוקיות לפי טעמים.

<u>מבנה הפסוקיות:</u>

- כל המקורות מציגים ממוצע שנע בין 1.27 ל-1.7, גם פה אין פערים מהותיים בין המקורות אך הפערים גדולים יותר לעומת החלוקה לפי פסוקיות הטעמים.
- המקור D1 בעל הממוצע הגבוה ביותר של פסוקיות לפי מבנה הפסוקיות והמקור E בעל הממוצע הנמוך ביותר של פסוקיות לפי מבנה הפסוקיות.

<u>מסקנות</u>

המקור עם הממוצע הנמוך ביותר לפי חלוקה למבנה פסוקיות גבוה מהמקור עם הממוצע הגבוה ביותר לפי טעמים אך במספרים נמוכים. דבר היכול להעיד על מבנה משפטים טיפה יותר מורכב.

אין עקביות במקורות מבחינת סדר הנתונים הממוין בין שתי השיטות.

עומקי פסוקים

ממוצע של עומקי הפסוקים לפי מקור:

על פי עץ הטעמים		על פי עץ תלויות	
D2	2.230769	D2	6.903846
Dn	2.167155	D1	6.321429
E	2.159451	Dn	6.228739
D1	2.154334	J	6.090310
J	2.147695	P	6.035673
P	2.088258	E	6.004211
O	2.077922	O	5.831169
R	2.066202	R	5.735192

<u>השוואה בין עץ הטעמים לעץ התלויות:</u>

ממוצע עומקי הפסוקים לפי עץ הטעמים נע בין 2.06 ל-2.23.

ממוצע עומקי הפסוקים לפי עץ התלויות נע בין 5.73 ל-6.9, ערכים גבוהים משמעותית מעומקי הטעמים.

עץ התלויות חושף מורכבות תחבירית עמוקה יותר לעומת עץ הטעמים, כפי שניתן לראות מהפער הגדול בין הממוצעים.

<u>דפוסים בין המקורות:</u>

מקור D2 מציג את העומק הממוצע הגבוה ביותר בשתי השיטות.

מקור R מציג את העומק הממוצע הנמוך ביותר בשתי השיטות.

עבור המקורות הנותרים, אין עקביות מבחינת סדר הנתונים הממוין בין שתי השיטות.

<u>מסקנות</u>

כל המקורות מציגים ממוצע עומק פסוקים דומה מאוד בעץ טעמים (~2) ובעץ תלויות (5-6), ללא הבדל משמעותי ביניהם.

ממוצע עומק הפסוקים אינו כלי מובהק להפרדת המקורות, כיוון שהוא זהה בכל המקורות שנבדקו.

הפסוקיות ושכיחותן

<u>תדירות הפונקציות הדקדוקיות במקורות השונים</u>

המקור J מאופיין בתחביר עשיר ומגוון, עם ריבוי פעלים (Pred), נושאים (Subj), ושימוש גבוה יחסית בשאלות (Ques). הדבר מעיד על אופי נרטיבי עם דיאלוגים וסגנון לשוני דינמי.

המקור P מציג שימוש נמוך בשאלות (Ques) ובמבנים רגשיים (Intj, Exst), דבר שמחזק את אופיו המשפטי-הלכתי ואת הסגנון הישיר שבו משתמשים.

המקור E כולל מספר רב של משפטים מחוברים (Conj) ופעלים (Pred), מה שמעיד על מבנים לשוניים מפורטים עם תיאורים רבים.

המקורות D1 ו-D2 מציגים מבנים לשוניים פשוטים יחסית, עם פחות משפטים מחוברים (Conj) ומורכבים, מה שעשוי להעיד על סגנון יותר ישיר או תמציתי.

המקור Dn שומר על מבנה מאוזן, עם שילוב של חיבורים ושמות עצם, מה שיכול להצביע על טקסטים רטוריים אך גם תיאוריים.

המקור R כולל תדירות נמוכה יחסית של פונקציות דקדוקיות מורכבות, דבר שיכול להעיד על סגנון מאורגן או מסכם של טקסטים.

המקור O מכיל שילוב של מבנים נרטיביים וחקיקתיים, עם מספר ממוצע של פעלים וחיבורים.

<u>תדירות סוגי הביטויים</u>

המקורות J ו-E מובילים בשימוש בביטויי פועל (VP) ושם עצם (NP), מה שמעיד על טקסטים עם דגש חזק על פעולות, דמויות ואירועים.

המקור P משתמש פחות בביטויי פועל ויותר בביטויי יחס (PP), מה שמדגיש את אופיו החקיקתי והפורמלי.

D1 ו-D2 כוללים ביטויי יחס (PP) רבים יחסית, דבר שיכול להצביע על דיוק בתיאור יחסים לוגיים, זמניים ומרחביים.

Dn מציג תמהיל מגוון בין סוגי הביטויים, עם נוכחות של ביטויי פועל ושמות עצם אך ללא דומיננטיות ברורה של אחד מהם.

O ו-R מציגים שימוש נמוך יותר בביטויי פועל אך מבנים מורכבים עם ביטויי יחס, דבר שיכול להעיד על טקסטים פחות נרטיביים ויותר הצהרתיים.

השוואה בין המקורות – מגמות עיקריות

המקור J הוא המקור עם התחביר המגוון ביותר, והוא מתאפיין בנרטיב עשיר ושפה דינמית.

המקור P הוא הפורמלי ביותר, עם מבנים תחביריים פשוטים יותר, פחות פעלים ושימוש גבוה בביטויים מבניים.

המקור E מציג ריבוי של פעלים וחיבורים, מה שמעיד על מבנים תחביריים רחבים ומפורטים.

D1, D2, ו-Dn מכילים מבנים מאוזנים יותר, ללא נטייה חזקה לנרטיב או לחקיקה, אך עם שימוש משמעותי בביטויי יחס.

המקור R נראה כמקור מסכם או מסודר יותר, עם פחות מבנים תחביריים מסועפים.

O מציג שילוב בין תיאורים נרטיביים והגדרות פורמליות, ללא דומיננטיות ברורה של מבנה תחבירי אחד.

מילים ייחודיות לפי מקור

מקור	סך כל המילים במקור	מילים ייחודיות במקור	אחוזים
O	896	791	88.281250
D2	921	680	73.832790
Dn	5179	3212	62.019695
D1	7081	4137	58.423951
J	14426	8443	58.526272
E	13587	7528	55.405903
R	3201	1767	55.201500
P	34596	14323	41.400740

מסקנות

מקור O הוא בעל האחוז הגבוה ביותר של מילים ייחודיות ומקור P בעל האחוז הנמוך ביותר של מילים ייחודיות. ייתכן שנתונים אלה מתקשרים לכך שסך כל המילים במקור O הוא הנמוך ביותר ואילו במקור P סך כל המילים הוא הגבוה ביותר.

ביצענו גם ניתוח סטטיסטי על מילים ייחודיות לפי כל פסוק בספר, האחוזים שהתקבלו שאפו ל-100% בכל אחד מהספרים ולכן לא ניתן להסיק מסקנות רלוונטיות מהנתון הזה. ניתוח הנתונים לפי כל פסוק (משפט קצר יחסית) יכול לגרום לכל מילה להיחשב ייחודית.

אורך מילים במקורות

ממוצע של אורכי מילים לפי מקור:

R	8.388316
D2	8.241042
D1	8.124700
P	7.988669
O	7.940848
Dn	7.921220
J	7.916540
E	7.842791

<u>מסקנות</u>

קיים פער בין המקורות באורך המילים הממוצע, אך ההבדלים אינם חדים מאוד. למרות ההבדלים, כולם נעים סביב אורך של 7.8–8.4 אותיות, כלומר השפה המקראית שומרת על אחידות יחסית.

R מכיל את המילים הארוכות ביותר, ואילו E את הקצרות ביותר.

תבניות הפסוקים

ביצענו ניתוח סטטיסטי לפי תבניות הפסוקיות אך לא הגענו למסקנות רלוונטיות. בנוסף, נותחו מבני העצים על פי המילים וההקשרים שלהם, ולכן הוספת ייצוג נוסף ללא מילים עלולה להעמיס מידע נוסף מבלי לספק תובנות חדשות, מה שעלול להפוך את הניתוח לפחות ממוקד ויעיל ולכן בחרנו לא להכניס ניתוח זה לסיכום.

מסקנות סופיות - נתונים סטטיסטיים בפילוח לפי מקור

ברוב הניתוחים שבוצעו ניתן לראות שמקור D2 ממוקם בראש הרשימה.

מקור R כמעט תמיד בתחתית הרשימה, דבר המצביע על מבנים פשוטים יותר, כנראה בשל אופיו כעורך שמטרתו לקשר בין טקסטים שונים בצורה ברורה וללא מורכבויות מיותרות.

גם פה, בדומה לניתוח לפי חלוקה לספרים, ניתן לראות שבניתוח אורך מילים מקור R מופיע במיקום הגבוה ביותר בניגוד לשאר הניתוחים שם נמצא בתחתית הטבלה.

מודל המסווג

המודל שבנינו מבצע סיווג של פסוקים לפי חומש (או לפי מקור בהתאמה) בהתבסס על מאפיינים לשוניים כגון: מילים, שורשים,חלקי דיבר,עץ תחבירי ושילוב של כל המאפיינים יחד.

לאחר מכן, הוא מאמן מודלים של למידת מכונה (Random Forest ו- SVM) על כל מאפיין בנפרד, ואז מודד ביצועים של כל מאפיין.
בשלב הראשון נבצע הוצאת נתונים מתאימים מעצי התלותיות שלנו (קבצי ה-dicta) ולאחר מכן מקובץ teamim-trees.txt שמכיל בתוכו את עצי הטעמים של הפסוקים.
מקבצי ה-dicta אנחנו מוציאים שלושה מאפיינים מרכזיים מכל פסוק: המילים כפי שהן מופיעות בטקסט, צורת השורש של כל מילה והאם המילה היא שם עצם, פועל, תואר וכו'.
כלומר, לכל פסוק נשמרים המילים, השורשים, חלקי הדיבר והחומש אליו הוא שייך.

מקובץ ה-teamim-trees.txt אנחנו נוציא את המבני עץ תחביריים של הפסוקים. כלומר לכל פסוק משויך מבנה עץ תלותי שמיוצג כטקסט.

בשלב השני, ניצור ארבע רשימות נפרדות עבור כל פסוק, כל אחת עם סוג מידע אחר.
על כך רשימה נשתמש ב-TF-IDF כדי להמיר את המידע הטקסטואלי למספרים שהמודלים יכולים לעבד.
כלומר קיבלנו שלכל מאפיין יש וקטור מספרי שמייצג את הפסוקים.
בנוסף, ניצור את הווקטור המשולב כדי לבדוק האם שימוש בכל המאפיינים יחד משפר את הביצועים.

במהלך המודל נשתמש בשני סוגי מסווגים וב2 שיטות validation שונות:
למה בחרנו במודלים Random Forest ו-SVM?

1. <u>Random Forest</u> – מודל מבוסס עצי החלטה.
   יתרון: מתאים לבעיות עם הרבה תכונות (features) שונות.
   יתרון: לא דורש שינוי בקלט – יכול לעבוד ישירות על וקטורים של מילים או תכונות תחביריות.
   חיסרון: פחות טוב בבעיות עם הרבה קשרים מורכבים בין התכונות.
2. <u>SVM (Support Vector Machine)</u> – מודל שמתאים להפרדה של קבוצות שונות בטקסט.
   יתרון: טוב מאוד בהפרדה בין קטגוריות שונות, במיוחד אם יש הבחנה ברורה במבנה השפה.
   יתרון: עמיד בפני רעש, עובד טוב במיוחד על דאטה עם מספר גבוה של תכונות.
   חיסרון: יכול להיות איטי עם דאטה גדול מאוד.

ניסינו להשתמש בשניהם על מנת להשוות בין התוצאות ולראות איזה מודל עובד טוב יותר על נתונים מהסוג שלנו.
כדי להעריך עד כמה המודל שלנו מדויק, השתמשנו בשני סוגים של ולידציה (Validation):

1. <u>ולידציה צולבת (Cross-Validation - 10-Fold)</u>:
   בולידציה צולבת אנחנו מחלקים את הנתונים ל-10 קבוצות שונות:
   כל פעם 9 קבוצות משמשות לאימון והקבוצה הנותרת משמשת לבדיקה.
   התהליך חוזר 10 פעמים, כך שכל דוגמה משמשת גם לאימון וגם לבדיקה.

2. <u>חלוקה לסט אימון וסט בדיקה (Train-Test Split)</u>:
כדי לבדוק איך המודל פועל על נתונים שלא ראה, אנחנו מחלקים את הנתונים ל- 80% אימון ו- 20% בדיקה.

<u>תוצאות המסווג של החומשים</u>

לפי מילים:

	Random Forest			SVM		
	accuracy: 0.677			accuracy: 0.749		
	precision	recall	f1-score	precision	recall	f1-score
Deuteronomy	0.663	0.666	0.664	0.774	0.734	0.754
Exodus	0.671	0.574	0.619	0.682	0.657	0.669
Genesis	0.703	0.843	0.767	0.814	0.885	0.848
Leviticus	0.685	0.622	0.652	0.753	0.709	0.730
Numbers	0.647	0.620	0.633	0.707	0.713	0.710

לפי שורשים:

	Random Forest			SVM		
	accuracy: 0.655			accuracy: 0.72		
	precision	recall	f1-score	precision	recall	f1-score
Deuteronomy	0.685	0.625	0.653	0.701	0.708	0.704
Exodus	0.614	0.553	0.582	0.647	0.652	0.650
Genesis	0.645	0.853	0.734	0.776	0.837	0.805
Leviticus	0.699	0.581	0.634	0.766	0.668	0.714
Numbers	0.663	0.589	0.624	0.706	0.689	0.698

לפי חלקי דיבר:

	Random Forest			SVM		
	accuracy: 0.410			accuracy: 0.396		
	precision	recall	f1-score	precision	recall	f1-score

0.397	0.390	0.405	0.390	0.375	0.406	Deuteronomy
0.259	0.239	0.282	0.285	0.247	0.337	Exodus
0.477	0.547	0.423	0.486	0.628	0.396	Genesis
0.337	0.331	0.343	0.340	0.308	0.381	Leviticus
0.445	0.410	0.486	0.459	0.399	0.542	Numbers

לפי מאפיינים תחביריים:

SVM			Random Forest			
accuracy: 0.269			accuracy: 0.270			
f1-score	recall	precision	f1-score	recall	precision	
0.114	0.078	0.214	0.206	0.177	0.248	Deuteronomy
0.161	0.128	0.219	0.180	0.152	0.220	Exodus
0.401	0.719	0.277	0.365	0.517	0.282	Genesis
0.010	0.005	0.066	0.079	0.052	0.166	Leviticus
0.230	0.182	0.313	0.307	0.302	0.313	Numbers

לפי כל המאפיינים ביחד:

SVM			Random Forest			
accuracy: 0.725			accuracy: 0.626			
f1-score	recall	precision	f1-score	recall	precision	
0.743	0.760	0.726	0.6	0.546	0.664	Deuteronomy
0.630	0.615	0.645	0.539	0.508	0.574	Exodus
0.808	0.837	0.781	0.705	0.859	0.598	Genesis
0.728	0.686	0.776	0.600	0.546	0.666	Leviticus
0.693	0.693	0.693	0.623	0.573	0.682	Numbers

לפי שילוב של המאפיינים מילים ושורשים:

	Random Forest			SVM		
	accuracy: 0.677			accuracy: 0.748		
	precision	recall	f1-score	precision	recall	f1-score
Deuteronomy	0.690	0.661	0.675	0.755	0.770	0.762
Exodus	0.622	0.545	0.581	0.676	0.640	0.658
Genesis	0.674	0.876	0.762	0.806	0.869	0.836
Leviticus	0.724	0.656	0.689	0.772	0.75	0.761
Numbers	0.690	0.589	0.635	0.713	0.686	0.699

ניתוח תוצאות המסווג לפי חומשים:

כאשר המסווג משתמש במילים כמאפיין יחיד, מתקבלת רמת דיוק גבוהה יחסית, זו רמת הדיוק הגבוהה ביותר מבין המאפיינים. בנוסף, כאשר בשימוש בשורשים כמאפיין יחיד, מתקבלת גם כן רמת דיוק גבוהה יחסית, קרובה מאוד לרמת הדיוק של השימוש במילים.

מאפיינים תחביריים הם המאפיין עם רמת הדיוק הנמוכה ביותר לסיווג חומשים בשני המודלים.

השפעת המאפיינים התחביריים (מבנה עץ הטעמים) וחלקי הדיבר POS פחותה ביחס למילים ושורשים - התקבלו ביצועים שאינם מספקים מבחינת רמות הדיוק. לכן ניתן להניח שחלקי הדיבר כמאפיין ומבני העץ כמאפיין אינם משפרים את יכולות הסיווג שלנו ולא מהווים מאפיין חזק ומשמעותי לסיווג הפסוקים.

בתהליך העבודה שלנו, רצינו לבדוק את רמת הדיוק במסווג על כל המאפיינים יחד, השילוב הניב ביצועים טובים, והדיוק של שני המודלים מתקרב לדיוק של הסיווג לפי מילים, אך עדיין נמוך ממנו.

בנוסף, משום שהשימוש במסווג לפי מילים ושורשים בנפרד הניב את הדיוקים הגבוהים מבין כל מאפיין בנפרד, בדקנו אפשרות של שימוש במסווג לפי שני המאפיינים ביחד. ראינו כי עבור RF הדיוק הוא 0.677 ועבור SVM הדיוק הוא 0.748 - כמעט זהה לדיוק של שימוש במילים כמאפיין בלבד.

SVM הוא המודל שמספק את התוצאות הטובות ביותר בסיווג לפי מילים ושורשים וכל המאפיינים ביחד, אך גם הוא לא מצליח להוציא ביצועים טובים בסיווג לפי חלקי דיבר ומאפיינים תחביריים.

RF פחות טוב מ-SVM במרבית המקרים, אך בסיווג לפי חלקי דיבר מספק תוצאות סבירות יותר.

תוצאות המסווג של המקורות

לפי מילים:

SVM			Random Forest			
accuracy: 0.719			accuracy: 0.657			
f1-score	recall	precision	f1-score	recall	precision	
0.651	0.6	0.712	0.513	0.410	0.684	D1
0	0	1	0	0	1	D2
0.4	0.308	0.567	0.336	0.25	0.515	Dn
0.565	0.536	0.596	0.480	0.384	0.640	E
0.656	0.685	0.629	0.566	0.549	0.585	J
0	0	0	0.125	0.066	1	O
0.848	0.923	0.785	0.784	0.942	0.672	P
0.723	0.586	0.944	0.673	0.534	0.911	R

לפי שורשים:

SVM			Random Forest			
accuracy: 0.725			accuracy: 0.655			
f1-score	recall	precision	f1-score	recall	precision	
0.674	0.642	0.709	0.377	0.242	0.851	D1
0	0	1	0	0	1	D2
0.392	0.294	0.588	0.229	0.147	0.526	Dn
0.576	0.536	0.621	0.523	0.447	0.629	E
0.668	0.723	0.620	0.569	0.577	0.561	J
0	0	0	0	0	1	O
0.856	0.925	0.797	0.790	0.957	0.672	P
0.644	0.5	0.906	0.627	0.465	0.964	R

לפי חלקי דיבר:

SVM			Random Forest			
accuracy: 0.491			accuracy: 0.482			
f1-score	recall	precision	f1-score	recall	precision	
0.073	0.042	0.285	0.055556	0.031	0.230	D1
0	0	1	0	0	1	D2
0	0	0	0.123457	0.073	0.384	Dn
0.26	0.205	0.354	0.28483	0.242	0.345	E
0.340	0.286	0.420	0.226866	0.178	0.311	J
0	0	1	0	0	1	O
0.664	0.902	0.526	0.656	0.873	0.525	P
0	0	1	0.425	0.293	0.772	R

לפי מאפיינים תחביריים:

SVM			Random Forest			
accuracy: 0.445			accuracy: 0.416			
f1-score	recall	precision	f1-score	recall	precision	
0	0	1	0.049	0.031	0.111	D1
0	0	1	0	0	1	D2
0	0	1	0.047	0.029	0.125	Dn
0	0	0	0.078	0.052	0.153	E
0.009	0.004	0.5	0.126	0.084	0.253	J
0	0	1	0	0	0	O
0.616	0.998	0.446	0.599	0.865	0.458	P
0	0	1	0.093	0.051	0.5	R

לפי כל המאפיינים ביחד:

SVM			Random Forest			
accuracy: 0.719			accuracy: 0.596			
f1-score	recall	precision	f1-score	recall	precision	
0.630	0.610	0.651	0.308	0.2	0.678	D1
0	0	1	0	0	1	D2
0.429	0.338	0.589	0.183	0.117	0.421	Dn
0.588	0.526	0.666	0.361	0.263	0.574	E
0.641	0.680	0.606	0.486	0.422	0.573	J
0	0	0	0	0	1	O
0.850	0.931	0.782	0.734	0.967	0.592	P
0.688	0.534	0.968	0.611	0.448	0.962	R

לפי שילוב של המאפיינים מילים ושורשים:

SVM			Random Forest			
accuracy: 0.726			accuracy: 0.661			
f1-score	recall	precision	f1-score	recall	precision	
0.659	0.621	0.702	0.425	0.315	0.652	D1
0	0	1	0	0	1	D2
0.450	0.367	0.581	0.231	0.161	0.407	Dn
0.569	0.531	0.612	0.501	0.415	0.632	E
0.659	0.699	0.623	0.585	0.605	0.565	J
0	0	0	0	0	0	O
0.860	0.931	0.800	0.805	0.950	0.698	P
0.688	0.534	0.968	0.652	0.517	0.882	R

ניתוח תוצאות המסווג לפי מקורות:

כאשר המסווג משתמש במילים או בלמות (שורשים) בתור מאפיין יחיד, התקבלו ביצועים די דומים ברמות הדיוק - הטובים ביותר עבור סיווג מקור הפסוק. כלומר, המילים והשורשים משמשים כמאפיין מרכזי ומאפשרות זיהוי טוב של מקור הפסוק.

מאפיינים תחביריים הם המאפיין עם רמת הדיוק הנמוכה ביותר לסיווג חומשים בשני המודלים.

השפעת המאפיינים התחביריים (מבנה עץ הטעמים) וחלקי הדיבר POS פחותה ביחס למילים ושורשים - התקבלו ביצועים שאינם מספקים מבחינת רמות הדיוק. לכן ניתן להניח שחלקי הדיבר כמאפיין ומבני העץ כמאפיין אינם משפרים את יכולות הסיווג שלנו ולא מהווים מאפיין חזק ומשמעותי לסיווג הפסוקים.

בתהליך העבודה שלנו, רצינו לבדוק את רמת הדיוק במסווג על כל המאפיינים יחד, השילוב הניב ביצועים טובים (במיוחד בשימוש עם מודל ה-SVM) אך עדיין ללא כל שיפור משמעותי לעומת השימוש במילים או למות (שורשים) כמאפיין יחיד בסיווג.

בנוסף, משום שהשימוש במסווג לפי מילים ושורשים בנפרד הניב את הדיוקים הגבוהים מבין כל מאפיין בנפרד, בדקנו אפשרות של שימוש במסווג לפי שני המאפיינים ביחד. השילוב ביניהם הראה שיפור בביצועים בהשוואה לשימוש במאפיין אחד בלבד, והציג גם ביצועים טובים יותר מהשילוב של המאפיינים יחד, כלומר שיפר את היכולת של המסווג לזהות את מקור הפסוק בצורה טובה יותר. שילוב שני המאפיינים מניב את הדיוק הטוב ביותר עבור מסווג זה.

נקודה נוספת שיש להתייחס אליה היא הקושי באימון וסיווג קטגוריות בעלות כמות נמוכה של דוגמאות. במקורות כמו D2 ו-O, שבהם כמות הדוגמאות הייתה נמוכה, ערכי המדדים היו בקיצון (0 ו-1) במרבית המאפיינים. דבר זה מצביע על הקושי של המודלים להתמודד עם קטגוריות בעלות דוגמאות מוגבלות ולהשיג תוצאות מדויקות. מודלים של למידת מכונה, כמו SVM ו-RF, דורשים כמות מסוימות של דוגמאות על מנת לייצר חוקים כללים וייצוג מלא של הקטגוריות. אחרת, המודל עלול לא להצליח לזהות את הדפוסים הכללים שמאפיינים את הקטגוריה כך שעלול להיווצר מצב של overfitting או underfitting ובעיות נוספות - ממש כמו במקרה שלנו.

SVM הוא המודל שמספק את התוצאות הטובות ביותר בסיווג לפי כל מאפיין לבד, כל המאפיינים ביחד, ובשילוב של מילים ושורשים.

סיכום ומסקנות

בבואנו לבצע את המחקר, ביקשנו לחקור את השינויים המתרחשים בין ספרי התורה השונים ובין המקורות השונים לפי השערת התעודות.
מטרת המחקר הייתה להבין, בעזרת כלים סטטיסטיים וניתוחים כמותיים, אם ניתן לזהות הבדל משמעותי במאפיינים תחביריים ולשוניים בין החומשים ובין המקורות השונים.
עלו מספר תובנות חשובות שניתן להפיק מהמחקר שערכנו:

המאפיינים החזקים ביותר: עבור שני המסווגים, שימוש במאפיין המילים ומאפיין הלמות (שורשים) בנפרד הניב את התוצאות המדויקות ביותר של שני המסווגים. כלומר, הם המאפיינים החזקים בתוך קבוצת המאפיינים שבחנו ולהם יש את ההשפעה הגדולה ביותר על דיוק הסיווג פסוקים.

שוני לא מובהק: קיים שוני בין רמות הדיוק של המסווגים כאשר השתמשנו בשילוב של כל המאפיינים לעומת שילוב המאפיינים החזקים (מילים ושורשים), אך השוני הזה אינו משמעותי. זאת אומרת, אפילו שהיו הבדלים ברמות הדיוק, הם לא היו גורמים מכריעים ויש סיבות נוספות שעשויות להסביר את השוני הזה, ולכן אין להסיק מהם מסקנה חזקה לעתיד.

מודלים ושיטות סיווג: מודל ה-SVM הפיק תוצאות טובות ביותר במרבית הפעמים בהשוואה למודל RF. בנוסף, בדקנו מודלים נוספים ובניהם אלגוריתם רשת הנוירונים - אלגוריתם זה נמצא בשימוש נרחב מאוד בשנים האחרונות בתחום ה-Deep Learning, ולכן חשבנו שיכול להועיל בתהליך הסיווג שלנו. אבל, תוצאותיו היו נמוכות יותר בדיוקן לעומת SVM ולכן בחרנו לא לפרט עליו בדוח זה.

שיפור מדדים ושימוש באחוזי הולידציה: ההתאמה של אחוזי ההולידציה עזרה להשיג תוצאות אופטימליות, כלומר ישנה חשיבות לא רק במאפיינים עצמם אלא גם בהגדרת הפרמטרים המתאימים בתהליך הסיווג.

מסקנה כללית: ניתוח הנתונים הסטטיסטיים מצביע על כך שבמרבית המאפיינים שנבחנו, אין שוני משמעותי בין החומשים או המקורות השונים מבחינה תחבירית ולשונית. נבחין כי המסווגים הגיעו לדיוק גבוה, אך לא מושלם. הדבר תומך בתפיסה כי אין שוני מהותי בין ספרי התורה השונים ומקורותיהם, לפחות לא במאפיינים שנבחנו במחקר זה.

צעדים להמשך: בכדי להמשיך ולהעמיק את המחקר עם תובנות נוספות, ניתן לשקול מספר צעדים להמשך שחשבנו שיכולים להיות רלוונטים.
ביניהם: שימוש בשיטות ניתוח נוספות במידה וקיימות, הרחבת המאפיינים שנבדקו (כגון: אורך פסוקיות, תדירות מילים מסויימות, שימוש במילים נדירות וכדומה), בחינת אלגוריתמים נוספים בתחום ה-Deep Learning שאולי יכולים להפיק תוצאות מדויקות יותר ובנוסף הרחבת בסיס הנתונים על ידי הכללת מקורות נוספים אם הדבר אכן אפשרי.

קצת על הקוד:
לאורך כל העבודה, כתבנו קבצי קוד לשליפת נתונים (הפסוקים), ייצוגם על פי שיטות הניתוח השונות, חלוקתם על פי חומשי התורה/מקורות תעודה ולבסוף הרכבנו את המסווגים אשר מתבססים על פלטי הקוד שכתבנו לאורך הפרויקט.
את הקוד ניסינו לממש בצורה המדויקת ביותר ובאופן מודולרי וקל לשימוש.
כמו כן, ביצענו שימוש במגוון רחב של ספריות המיועדות לעיבוד נתונים, למידת מכונה, עיבוד שפה טבעית, ניתוח טקסטים,ניתוח עצים, עבודה עם XML ועוד. תוך שימוש בכלים מתקדמים כמו

transformers, sklearn, PyTorch ועוד, ספריות המיועדות לעבודה עם טקסטים, נתונים ומודלים של למידת מכונה.
צירפנו להגשה קובץ ZIP עם כל קבצי הקוד שלנו מהפרויקט.