

תכנית עבודה וארכיטקטורת פלטים

1. תיאור הדאטה והקבצים (Corpora & JSON)

1.1. המאגרים (קורפוסים) העיקריים

המחקר משווה בין שלוש קטגוריות על (Main Corpora):

1. **Biblical Hebrew (BH)**: לשון המקרא.
2. **Rabbinic Hebrew (RH)**: לשון חכמים/חז"ל.
3. **Modern Hebrew (MH)**: עברית ישראלית בת זמננו.

1.2. תתי-מאגרים (Sub-Corpora)

כל קורפוס ראשי מחולק לתתי-ז'אנרים כדי לאפשר רזולוציה עדינה יותר של הניתוח:

- **היסטורי**: תנ"ך, משנה (שישה סדרים), רמב"ם (משנה תורה).
- **מודרני**: חדשות, ספרות (מקור ומתורגמת), בלוגים, פורומים (תפוז), טקסטים רפואיים/מדעים.

1.3. פורמט הנתונים (Dicta JSON)

הנתונים מגיעים בקבצי JSON שעברו ניתוח מורפולוגי ותחבירי על ידי הכלים של Dicta.

כל משפט מיוצג כאובייקט המכיל רשימת טוקנים (tokens). לכל טוקן (מילה) יש מידע עשיר הכולל:

- **Lex**: הלמה (צורת הבסיס המילונית).
- **Morph**: חלק דיבר (POS) ומאפייני הטיה (מין, מספר, גוף, זמן, תחיליות) ¹.
- **Syntax**: עץ תלויות – הצבעה למילת הראש (head) ותפקיד תחבירי (relation) כמו נושא, מושא וכו' ².

2. שאלות מחקר ומדדים סטטיסטיים

שאלה 1: מאפיינים סטטיסטיים כלליים ומורכבות המשפט

רציונל (בהקשר לדורון): דורון מערערת על התפיסה שלשון המקרא היא "פשוטה" ושרשרתית (Paratactic). היא טוענת שהתחביר המקראי מורכב ועמוק (Hypotactic), בדומה לעברית המודרנית ³. השוואה סטטיסטית של עומק העץ ואורך המשפט תאפשר לבדוק האם המודרנית אכן מציגה מורכבות הדומה למקרא, בניגוד לטענה המקובלת על פשטות המקרא.

מדדים:

- אורך משפט ממוצע (מספר טוקנים).
- עומק מילולי (יחס בין מספר הלמות הייחודיות למספר הטוקנים).
- עומק העץ התחבירי הממוצע.
- כיוון דרישות הקורס: נתונים סטטיסטיים כלליים, אורכי משפטים, עומק עצים.

שאלה 2: שימוש במילות שעבוד (Subordination)

רציונל (בהקשר לדורון): דורון מתארת תהליך של "התכנסות" (Convergence) בעברית המודרנית: אימוץ השלד התחבירי המקראי אך "הלבשתו" במילים חז"ליות⁴. ספציפית, היא מציינת כי המודרנית משתמשת במילת השעבוד "ש-" (מקור חז"לי) בתוך מבני זיקה שהם תחבירית מקראיים (במקום "אשר")⁵. בדיקת השכיחות תראה את המתח הזה בין הלקסיקון לתחביר.

מדדים:

- שכיחות יחסית של המילים: ש-, אשר, כי, כאשר, מאשר, אם, פן⁶.
כיסוי דרישות הקורס: שכיחות מילות/מילות השעבוד.

שאלה 3: סדר מילים תחבירי (V1 vs V2)

רציונל (בהקשר לדורון): דורון טוענת כי התחביר המודרני אימץ מחדש מאפיינים מקראיים. מאחר והבחנה מרכזית בין המקרא (שנוטה ל-V1, פועל לפני נושא) לחז"ל והמודרנית (שנוטים ל-V2) היא קריטית, הבדיקה תבחן האם למרות הנטייה הכללית ל-SVO, קיימים במודרנית מבני V1 ("התחיל הגשם") בשיעור המקרב אותה למקרא יותר מאשר לחז"ל, כחלק מהטענה ל"דילוג ההיסטורי" 7777.

מדדים:

- אחוז המשפטים בהם הפועל מקדים את הנושא (V1).
- אחוז המשפטים בהם הנושא מקדים את הפועל (V2).
- כיסוי דרישות הקורס: מבני V1 ומבני V2, סוג הקשרים התחביריים.

שאלה 4: הבעת שייכות (Possession)

רציונל (בהקשר לדורון): גם כאן דורון מדגימה את עקרון ה"התכנסות": המודרנית משתמשת במילה "של" (שמקומה בלשון חז"ל), אך לעיתים בתוך מבנים תחביריים מורכבים יותר המזכירים את הסמיכות המקראית או שילוב שלהם⁸. הבדיקה תשווה את היחס בין סמיכות חבורה (מקראית) לבין השימוש ב"של" (חז"לי) כדי לכמת את המיקום של המודרנית על הרצף הזה.

מדדים:

- שכיחות מבני סמיכות (Construct-state).
- שכיחות שימוש במילת היחס "של" (Prepositional possessive).
- יחס (Ratio) בין סמיכויות ל"של".
- כיסוי דרישות הקורס: מבני Prepositional possessives ומבני Construct-state possessive.

שאלה 5: צורות המקור (Gerund vs Infinitive)

רציונל (בהקשר לדורון): זוהי הראיה החזקה ביותר של דורון ("הנקודה הקריטית")⁹. היא טוענת שלשון חז"ל איבדה את ה-Gerund המקראי (שם פועל עם נושא, כמו "בצאת ישראל"), ואילו העברית המודרנית "החיותה" אותו מחדש (למשל: "עם בואו")¹⁰. בדיקה זו נועדה לאשש ישירות את טענת ה"דילוג" על פני לשון חז"ל בתחום זה.

מדדים:

- שכיחות Gerund: שם פועל עם נושא/כינוי קניין חבור (למשל: "בבואו").

- שכיחות Infinitive: שם פועל ללא נושא (למשל: "רצה ללכת").
- כיסוי דרישות הקורס: מבני Gerund ומבני Infinitive 11111111.

שאלה 6: פרופיל מורפולוגי ולקסיקלי

רציונל (בהקשר לדורון): דורון מצביעה על עיקרון של "מסומן" (Marked) מול "לא-מסומן" (Unmarked). מילים מקראיות ("עץ", "אף") משמשות במודרנית כברירת מחדל ניטרלית, בעוד מילים חז"ליות מקבילות ("אילן", "חוטם") נתפסות כמשלב גבוה או ספרותי 12. ניתוח התדירויות של זוגות אלו יבדוק האם הלקסיקון המודרני הבסיסי אכן מושתת על הרובד המקראי.

מדדים:

- התפלגות חלקי דיבר (POS).
- התפלגות זמנים (עבר/הווה/עתיד) 13131313.
- שכיחות זוגות מילים מהטבלה של דורון (למשל: עץ/אילן, שמש/חמה).
- כיסוי דרישות הקורס: קטגוריות מילים, מאפייני הטיה, טבלת המילים בעמוד 4.

3. תהליך העבודה (Pipeline)

3.1 שלב 1 – טעינת הקורפוסים וניתוח מקדים

טעינת כל קבצי ה-JSON לזיכרון. בשלב זה כל משפט מקבל תיוג של המקור שלו (main_corpus) והז'אנר הספציפי (sub_corpus).

3.2 שלב 2 – חילוץ פיצ'רים מרמת המשפט (Feature Extraction)

זהו השלב החשובי העיקרי. אנו מריצים פונקציה על כל משפט בודד ומחלצים ממנו וקטור של נתונים: אורכו, המילים הספציפיות שבו, המבנה התחבירי שלו וכו'. התוצר של שלב זה הוא "טבלת המאסטר" (ראה סעיף 4.1).

3.3 שלב 3 – סטטיסטיקה תיאורית והשוואות

ביצוע אגרגציה (Grouping) של הנתונים לפי קורפוסים. חישוב ממוצעים וסטיות תקן לכל אחד מהמדדים שהוגדרו, ויצירת טבלאות השוואה בין התקופות.

3.4 שלב 4 – מדידת מרחקים וסיווג

- חישוב מרחקים (Euclidean/Cosine) בין הוקטור הממוצע של העברית המודרנית לבין המקרא וחז"ל.
- אימון מודל סיווג (כגון Logistic Regression) על נתוני המקרא וחז"ל, והפעלתו על נתוני העברית המודרנית כדי לבדוק לאן הוא מסווג אותם 14141414.

3.5 שלב 5 – הכנת פלטים לדוח

שמירת התוצאות לקבצים מסודרים (CSV ו-TXT) שישמשו לכתיבת הדוח הסופי.

4. ארכיטקטורת הפלטים (Output Architecture)

מערכת הפלטים בנויה בארבע שכבות (Layers):

4.1 Layer 1 – Master Features Table

- **שם הקובץ:** `all_sentences_features.csv`
- **תיאור:** טבלת ענק המכילה שורה לכל משפט בפרויקט.
- **עמודות:** מזהה משפט, קורפוס, תת-קורפוס, וכל הפיצ'רים שחולצו (בוליאניים ומספריים) עבור אותו משפט.
- **חשיבות:** זהו בסיס הנתונים הגולמי המעובד, המאפשר כל ניתוח עתידי ללא צורך בעיבוד מחדש.

4.2 Layer 2 – קבצי סטטיסטיקה ואגרגציה

קבצים אלו מכילים את הסיכומים ברמת הקורפוס/תת-קורפוס ומשמשים ישירות ליצירת הגרפים והטבלאות בדוח.

- **`corpus_overview_stats.csv`:** סטטיסטיקה כללית (כמות משפטים, מילים, אורך ממוצע, עומק עץ).
- **`subordination_words_stats.csv`:** טבלת שכיחויות של מילות השעבוד (ש/אשר/כי וכו') בכל קורפוס.
- **`doron_lexical_pairs_stats.csv`:** השוואת שכיחויות של זוגות המילים המבחינות (כפי שמופיע בטבלת דורון).
- **`word_order_v1_v2_stats.csv`:** אחוז משפטי V1 לעומת V2 בכל קורפוס.
- **`gerund_infinitive_stats.csv`:** שכיחות ויחס בין צורות מקור (Gerund) למקור נטוי (Infinitive).
- **`pos_distribution_stats.csv`:** התפלגות חלקי הדיבר (אחוז פעלים, שמות עצם וכו').
- **`possession_constructions_stats.csv`:** השוואה בין שימוש בסמיכות לבין שימוש ב"של".
- **`corpus_distance_matrix.csv`:** מטריצת מרחקים מתמטית המראה את הקרבה בין כל זוג קורפוסים.

4.3 Layer 3 – דוגמאות איכותניות

- **קבצים:** `example_sentences_v1_v2.txt`, `example_sentences_possessive.txt`
- **תיאור:** קבצי טקסט המכילים דוגמאות למשפטים אמיתיים מתוך הקורפוסים שזוהו כבעלי תכונה מסוימת.
- **חשיבות:** מאפשרים בדיקת נאותות (Sanity Check) לאלגוריתם ומספקים דוגמאות יפות לשילוב בגוף העבודה.

4.4 Layer 4 – תוצאות מסווגים

- **`classifier_results_historical.csv`:** מדדי הביצועים (דיוק, F1) של המודל בהבחנה בין מקרא לחז"ל.
- **`classifier_results_modern.csv`:** תוצאות הסיווג של המודרנית – איזה אחוז מכל תת-קורפוס מודרני סווג כ"מקראי" ואיזה כ"חז"לי".

5. מיפוי לקובצי הדוח (כיצד נשתמש בפלטים)

חלק בדוח הסופי	קבצי הפלט הרלוונטיים
מבוא ותיאור הנתונים	corpus_overview_stats.csv (טבלת היקף הנתונים).
שיטה (Method)	תיאור תהליך ה-Feature Extraction מתוך מבנה ה-all_sentences_features.csv.
תוצאות – סטטיסטיקה תיאורית	pos_distribution_stats.csv, subordination_words_stats.csv, doron_lexical_pairs_stats.csv
תוצאות – ניתוח תחבירי (ליבת המחקר)	word_order_v1_v2_stats.csv, gerund_infinitive_stats.csv, possession_constructions_stats.csv. בשילוב דוגמאות מ-Layer 3.
תוצאות – סיווג ומרחקים	corpus_distance_matrix.csv, classifier_results_modern.csv
דיון ומסקנות	אינטגרציה של כלל הממצאים לאישוש/הפרכת הטענה של דורון.