

תכנית עבודה ארכיטקטורת פלטימ

1. תיאור הדטה 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/ח'ל'.
- 3. Modern Hebrew (MH): עברית ישראלית בת זמננו.

1.2 תת-מאגרים (Sub-Corpora)

כל קורפוס ראשי מחולק לתת-ז'אנרים כדי לאפשר רזרולציה עדינה יותר של הניתוח:

- היסטורי: תנ"ר, משנה (שישה סדרים), רמב"ם (משנה תורה).
- מודרני: חדשות, ספרות (מקור ומטורגם), בלוגים, פורומים (תפוז), טקסטים רפואיים/מדעיים.

1.3 פורמט הנתונים (Dicta JSON)

הנתונים מגאים בקבץ JSON שעבורו ניתוח מורפולוגי ותחבيري 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 דבר (POS) ומאפייני הטיה (מין, מספר, גוף, זמן, תחיליות)¹.
- Syntax: עצ תלויות – הצבעה למילת הראש (head) ותפקיד תחבירי (relation) כמו נושא, מושא וכו².

2. שאלות מחקר ומדדים סטטיסטיים

שאלה 1: מאפיינים סטטיסטיים כלליים ומורכבות המשפט

רצינול (בהקשר לדرون): דורון מערעתה על התפיסה שלשון המקרא היא " פשוטה" ושרשרתית (Paratactic). היא טעונה שהתחביר המקראי מורכב ועמוק (Hypotactic), בדומה לעברית המודרנית 3. השוווא סטטיסטי של העז והרך המשפט מאפשר לבדוק אם המודרנית אכן מציגה מורכבות הדומה למקרא, בኒיגוד לטענה המקובלת על פשוטות המקרא.

מדדים:

- אורך משפט ממוצע (מספר טוקנים).
 - עורך מילולי (יחס בין מספר הלמות הייחודיות למספר הטוקנים).
 - עומק העז התחבيري הממוצע.
- כיסוי דרישות הקורס: נתונים סטטיסטיים כלליים, אורכי משפטיים, עומק עצים.

שאלה 2: שימוש במילות שעבוד (Subordination)

רצינול (בקשר לדורון): דורון מתארת תהליך של "התכנסות" (Convergence) בעברית המודרנית: אימוץ השלד התחבירי המקראי אף "לבשו" במילים חז"ליות. ספציפית, היא מצינית כי המודרנית משתמשת במילת השעבוד "ש-" (מקור חז"ל) בתוך מבני זיקה לשם תחבירית מקראיים (במקום "אשר")⁵. בדיקת השכיחות תראה את המתה זהה בין הלקסיקון לתחביר.

מגדדים:

- שכיחות יחסית של המילים: ש-, אשר, כי, כאשר, מאשר, אם, פן 6.
כיסוי דרישות הקורס: שכיחות מילות/מילות השעבוד.

שאלה 3: סדר מילים תחבירי (V1 vs V2)

רצינול (בקשר לדורון): דורון טוענת כי התחביר המודרני אימץ מחדש מבניים מקראיים. לאחר והבינה מרכזית בין המקרא (שנוטה ל-V1, פועל לפני נושא) לחז"ל והמודרנית (שנוטה ל-V2) היא קритית, הבדיקה תבחן האם למחרת הנטיה הכללית ל-SVO, קיימים במודרנית מבני V1 ("התחיל הגשם") בשיעור המקובל אותה למקרא יותר מאשר לחז"ל, חלק מהטענה לדילוג ההיסטורי"⁷⁷⁷⁷.

מגדדים:

- 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)

רצינול (בקשר לדורון): גם כאן דורון מדגימה את עקרון ה"התכנסות": המודרנית משתמשת במילה "של" (שמקומה בלשון חז"ל), אף לעתים בתוך מבנים תחביריים מורכבים יותר המזכירים את הסמיוכות המקראית או שילוב שלהם.⁸ הבדיקה תשווה את היחס בין סמיוכות חבורה (מקראית) לבין השימוש ב"של" (חז"ל) כדי לכמת את המיקום של המודרנית על הרצף זהה.

מגדדים:

- שכיחות מבני סמיוכות (Construct-state).
- שכיחות שימוש במילת היחס "של" (Prepositional possessive).
- יחס (Ratio) בין סמיוכיות ל"של".

כיסוי דרישות הקורס: מבני Construct-state Prepositional possessives ומבני Ratio.

שאלה 5: צורות המקור (e.g. Gerund vs Infinitive)

רצינול (בקשר לדורון): זהה הראיה החזקה ביותר של דורון ("הנקודה הקритית")⁹. היא טוענת שהשzon חז"ל איבדה את ה-Gerund המקראי (שם פועל עם נושא, כמו "בצאת ישראל"), ואילו העברית המודרנית "החייבת" אותו מחדש (למשל: "עם בואו")¹⁰. בדיקה זו נועדה לאשש ישירות את טענת הדילוג על פני לשון חז"ל בתחום זה.

מגדדים:

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

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

שאלה 6: פרופיל מורפולוגי ולקסיקלי

רצינול (בקשר לדرون): דرون מצביעה על עירון של "מוסומן" (Marked) מול "לא-מוסומן" (Unmarked). מילים מקריאות ("עז", "אף") משמשות במודרנית כברירת מחדל ניטרלית, בעוד מילים ח"לויות מקבילות ("אלן", "חוטם") נתפסות כמשלב גבוה או ספורתי 12. ניתוח התדריות של זוגות אלו יבדוק האם הלקסיקון המודרני הבסיסי אכן מושתת על הרובד המקרי.

מددים:

- התפלגות חלקית דבר (POS).
- התפלגות זמנים (עבר/הווה/עתיד) 13131313
- שכיחות זוגות מילים מהטבלה של דرون (למשל: עז/אלן, שמש/חמה).
- **כיסוי דרישות הקורס:** קטגוריות מילים, מאפייני הטיה, טבלת המילים בעמוד 4.

3. תהליך העבודה (Pipeline)

3.1. שלב 1 – טיענת הקורפוסים וניתוח מקדים

טעינה כל קבצי ה-JSON לזכרון. 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 החישובי העיקרי. אנו מרכיבים פונקציה על כל משפט בודד ומחלצים ממנו וקטור של נתונים: אורכו, המילים הספציפיות שבו, המבנה התחרيري שלו וכו'. התוצר של שלב זה הוא "טבלת המאסטר" (ראה עייף 4.1).

3.3. שלב 3 – סטטיסטיקה תיאורית והשוואות

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

3.4. שלב 4 – מדידת מרחקים וסיווג

- חישוב מרחקים Euclidean/Cosine) בין הווקטור הממוצע של העברית המודרנית לבין המקרה וחז"ל.
- אימון מודל סיווג (כגון Chosion Regression) על נתונים המקרה וחז"ל, והפעלתו על נתונים העברית המודרנית כדי לבדוק לאן הוא מסוויג אותו 14141414 .

3.5. שלב 5 – הכנת פלטים לדוח

שמירת התוצאות לקבצים מסוודרים (CSV ו-TXT) שיישמשו כתיבת הדוח הסופי.

4. ארכיטקטורת הפלטים (Output Architecture)

מערכת הפלטים בניתה באربע שכבות (Layers):

Layer 1 – Master Features Table .4.1

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

Layer 2 .4.2 – קבצי סטטיסטיקה ואגרגציה

קובצים אלו מכילים את הסיכומים ברמת הקורפוס/תת-קורפוס ומשמשים לשירות ליצירת הגրפים והטבלאות בדוח.

- **קבצי סטטיסטיקה כללית**: corpus_overview_stats.csv (עמוק עז).
- **קבצי סטטיסטיקה של מילוט השबוד**: subordination_words_stats.csv (בכל קורפוס).
- **קבצי סטטיסטיקה של זוגות המילים המבוחנות**: doron_lexical_pairs_stats.csv (כפי שמופיע בטבלת דורון).
- **קבצי סטטיסטיקה אחוז משפטים**: word_order_v1_v2_stats.csv (לעומת V2 בכל קורפוס).
- **קבצי סטטיסטיקה של נקורים (Gerund)**: gerund_infinitive_stats.csv (למקור נתוי (Infinitive)).
- **קבצי סטטיסטיקה של התפלגות חלקי הדיבור**: pos_distribution_stats.csv (אחוז פעלים, שמות עצם וכו').
- **קבצי סטטיסטיקה של שימוש בסמכיות בין שמות עצם**: possession_constructions_stats.csv (השווואה בין שימוש בסמכיות לבין שימוש בשם).
- **קבצי מרחוקים מתמטית המראה את הקרבה בין כל זוג קורפוסים**: corpus_distance_matrix.csv.

Layer 3 .4.3 – דוגמאות אינטנסיות

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

Layer 4 .4.4 – תוצאות מסווגים

- **קובץ תוצאות מסווגים מודדי הביצועים**: classifier_results_historical.csv (Dijk, F1) של המודל בהבנה בין מקרא לחז"ל.
- **קובץ תוצאות מסווג של המודרנית – איזה אחוז מכל תת-קורפוס מודרני סוג "מקראי" ואיזה סוג "חז"לי"**: classifier_results_modern.csv.

5. מיפוי לקובץ הדוח (כיצד השתמש בפלטים)

קובץ הפלט הרלוונטיים	חלק בדוח הסופי
מבוא ותיאור הנתונים	מבוא ותיאור הנתונים
תיאור תהליך Feature Extraction מתוך מבנה .all_sentences_features.csv	שיטת (Method)
תוצאות – סטטיסטיקה תיאורית	תוצאות – סטטיסטיקה תיאורית
תוצאות – ניתוח תחבירי (ליבת המחבר)	תוצאות – ניתוח תחבירי (ליבת המחבר)
תוצאות – סיווג וمرחקרים	תוצאות – סיווג ומרחקרים
דיון ומסקנות	דיון ומסקנות