פרויקט מסכם-עיבוד שפה טבעית, בניין התפעל

יובל צירלר 318644010
עמית אריה רוזנקרנץ 324223882
אוריין מורדוך 312118938
אביאל קללאו 323504670

דו"ח מסכם

באופן כללי ניתן לומר שהעבודה הייתה מאורגנת בצורה מובנת מאוד. עיקר הקשיים שהיו לנו היו בתהליך ה-parsing, כלומר ההמרה של אוסף הדוגמאות לוקטורים, במיוחד עבור הדוגמאות מהתנ"ך. הסיבה הייתה המבנה של הקובץ-החבילה שממירה קבצי xml למבנה נתונים של עץ לא ידעה כיצד להתמודד עם ה-syntactic info ולכן היה צריך לנקוט בגישה מעט שונה כדי להשיג את המידע הנחוץ.

הקבצים המצורפים לעבודה זו הם-

הקוד של ההמרה לווקטורים ושל המסווגים, לפי שני החלונות-parse-window1.py,parse-window2.py שנבדקו

מבנה הווקטור-vector structure.txt

(y_vector הדוגמאות המתוייגות בעברית מודרנית (התיוג נמצא תחת-modern_hebrew.conllu.txt

(glinert,blau הדוגמאות המתוייגות מהתנ"ך (התיוג נמצא בתגיות-text.xml

שיטות-בניין התפעל.docx-מדריך התיוג.

אופן בחירת הדוגמאות

לא הייתה גישה מיוחדת מאחורי בחירת הדוגמאות. פשוט נבחרו הדוגמאות הראשונות מהקבצים שמכילות פועל בבניין התפעל.

אופן בניית הווקטורים

ייצגנו כל מילה כווקטור שורה באורך 78, כאשר קואורדינטות 0-29 מתייחסות לתפקיד התחבירי וקואורדינטות 30-78-לתפקיד המורפולוגי. המשמעות של כל קואורדינטה מתוארת בקובץ vector structure המצורף. באופן כזה ניתן היה לפצל בקלות את הלמידה למאפיינים הרצויים באמצעות slicing של מערך הדוגמאות לטורים הרלוונטיים.

בנוסף לבדיקה של המאפיינים בדקנו גם את השפעת החלון-הקובץ parse-window1 הוא עם חלון של מילה אחת ו-parse-window2 הוא עם חלון של שתי מילים. ההשערה שלנו הייתה שהגדלת החלון תגדיל את הדיוק, וכפי שניתן לראות מהתוצאות שלנו זה אכן קרה בפועל.

מדריך התיוג לפי שיטת גלינרט

אם ניתן לתרגם לאנגלית ולהשתמש במילה became/become-2

אם אפשר לנסח מחדש כ-"האיש הלביש **את עצמו**" וכו' (בהתאם לנושא ולשורש)-3

אם ניתן לנסח מחדש עם המילים "התחזה ל-"-6

אם אפשר להחליף לפועל בבניין אחר שחוזר על עצמו פעמיים, עם ו' החיבור ביניהם (למשל "הלך והלך" במקום התהלך)-7

אם הצורה הפעילה של הפועל היא מבניין פיעל-בדקו אם מבצע הפעולה מופיע. אם לא אז 1, אם כן אז 4.

אם מבצע הפעולה הוא ברבים, האם המשפט עדיין יהיה הגיוני אם נעביר אותו ליחיד? אם לא אז 5.

אם מופיעות המילים "עם X" אחרי הפועל, האם המשפט יהיה הגיוני אם נמחק אותן? אם לא אז 5.

אם לא מופיעות המילים "עם X", האם ניתן לנסח אותו מחדש כך שכן יהיו? אם כן אז 5.

אם אף אחד מהקודמים לא מתאים אז 8.

נשים לב: כל תיוג בשיטת גלינרט מוכל בתיוג כלשהו בקבוצת בלאו, לפי הטבלה הבאה:

גלינרט	בלאו
1	4
2	5
3	1
4	4
5	2
6	3
7	6
8	6

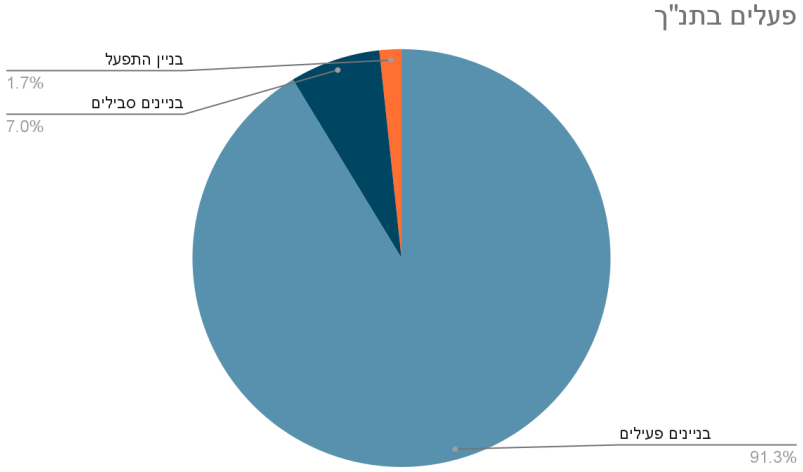
לכן מספיק לבנות תיוג לפי שיטת גלינרט בלבד, וממנו ניתן יהיה למצוא מיד את התיוג לפי שיטת בלאו.
מסיבה זו כל המידע הרשום בסטטיסטיקות מתייחס לצורת תיוג זו, והקוד יוצר כלל החלטה לשיטה זו.

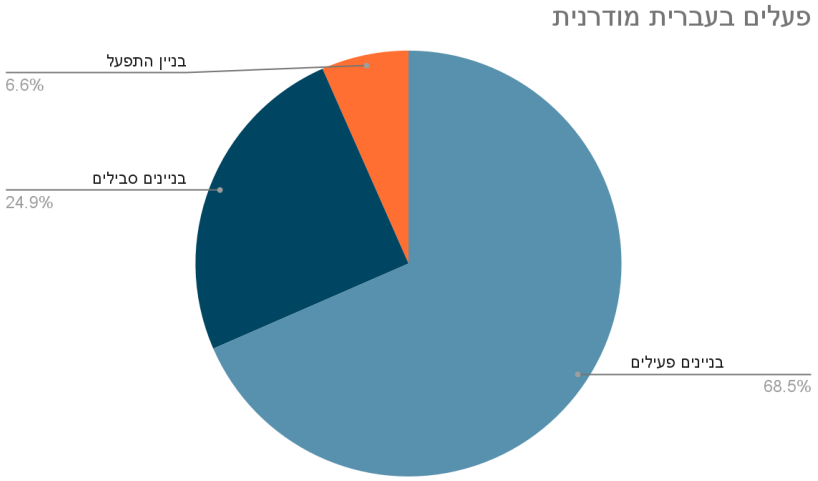
<u>סטטיסטיקות</u>

פעלים לפי בניין בכל מאגר הטקסטים:

בניין	פעלים בתנ"ך	פעלים בעברית מודרנית	סה"כ
פעל	48620	3923	52453
נפעל	4054	1538	5592
הפעיל	9167	1853	11020
הופעל	406	600	1006
פיעל	6727	1642	8369
פועל	488	563	1051
התפעל	1200	717	1917

ניתן לראות שבניין פעל הוא הנפוץ ביותר, ושהוא נפוץ בעיקר בתנ"ך. מנגד, חלקם של הפעלים הסבילים גדול הרבה יותר בעברית מודרנית.

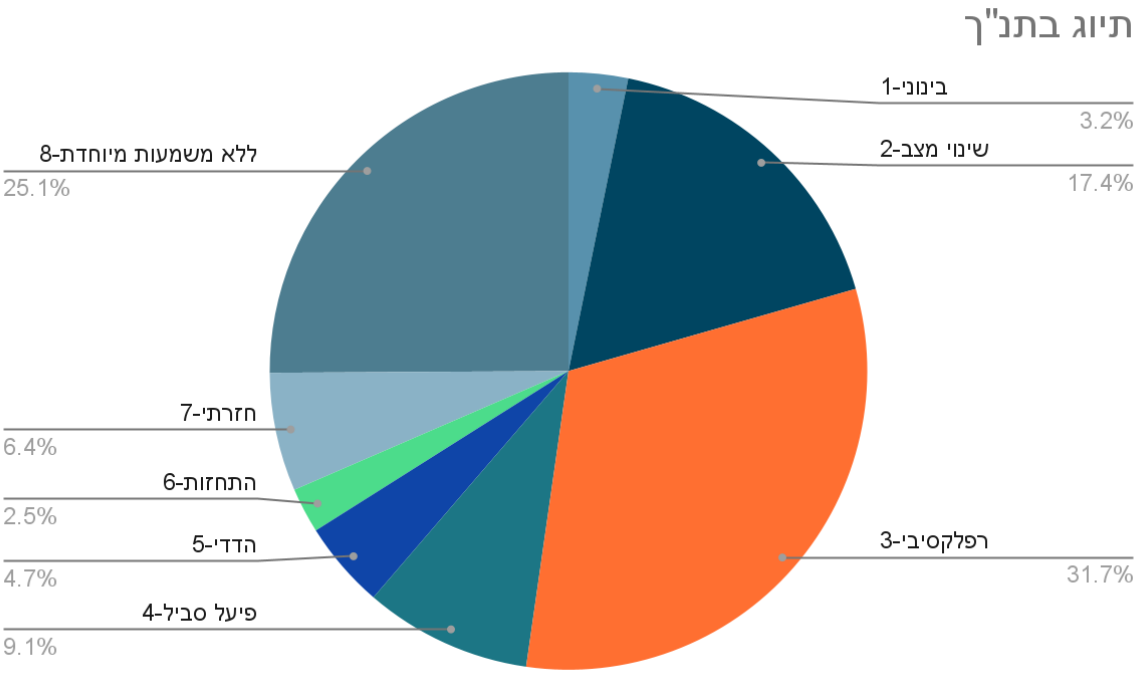


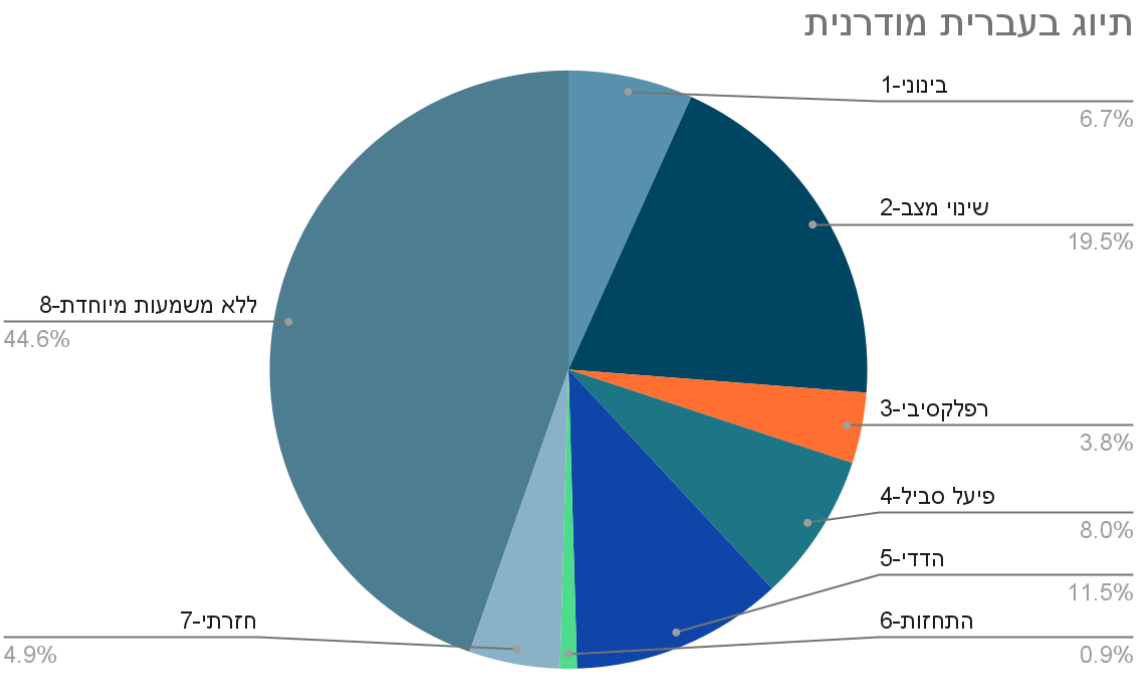


דוגמאות לפי התיוג בשיטת גלינרט, מבין הדוגמאות המתוייגות שלנו:

תיוג	תנ"ך	עברית מודרנית	סה"כ
1	17	37	54
2	92	107	199
3	168	21	189
4	48	44	92
5	25	63	88
6	13	5	18
7	34	27	61
8	133	245	378

ניתן לראות הבדלים בולטים בין התיוגים-בתנ"ך בולט במיוחד השימוש בבניין התפעל לפעלים רפלקסיביים (בעיקר עקב מספר מצומצם של פעלים שכיחים מאוד כמו "השתחווה"), בעוד שבעברית מודרנית קיימת שכיחות גבוהה הרבה יותר לפעלים הדדיים ולפעלים ללא משמעות מיוחדת לעומת התנ"ך.





<u>תוצאות</u>

המסווגים הטובים ביותר שמצאנו עבור כל אחת מהקבוצות מוצגים בטבלה:

<u>תנ"ך בלבד</u>

מאפיינים	גודל חלון	F Score	סוג מסווג
תפקיד מורפולוגי ותחבירי	1	0.32	SVM-RBF
תפקיד מורפולוגי	1	0.32	SVM-RBF
תפקיד תחבירי	1	0.34	SVM פולינומיאלי
תפקיד מורפולוגי ותחבירי	2	0.35	SVM-RBF
תפקיד מורפולוגי	2	0.35	SVM-RBF
תפקיד תחבירי	2	0.33	SVM-RBF

<u>עברית מודרנית בלבד</u>

מאפיינים	גודל חלון	F Score	סוג מסווג
תפקיד מורפולוגי ותחבירי	1	0.42	SVM לינארי
תפקיד מורפולוגי	1	0.43	SVM RBF או סיגמואיד
תפקיד תחבירי	1	0.43	SVM לינארי או סיגמואיד
תפקיד מורפולוגי ותחבירי	2	0.45	SVM לינארי
תפקיד מורפולוגי	2	0.43	SVM לינארי
תפקיד תחבירי	2	0.43	SVM RBF לינארי, פולינומיאלי או

<u>כל הדוגמאות</u>

מאפיינים	גודל חלון	F Score	סוג מסווג
תפקיד מורפולוגי ותחבירי	1	0.36	SVM RBF לינארי או
תפקיד מורפולוגי	1	0.37	SVM RBF
תפקיד תחבירי	1	0.36	SVM RBF או עץ החלטה אנטרופי
תפקיד מורפולוגי ותחבירי	2	0.38	SVM RBF
תפקיד מורפולוגי	2	0.38	SVM RBF
תפקיד תחבירי	2	0.36	SVM לינארי

<u>תובנות</u>

ניתן לראות עדיפות קלה לבדיקה בחלון של שתי מילים לפני ואחרי לעומת מילה אחת בלבד. הסיבה לכך ברורה-המילים שלפני ואחרי המילה נותנות הקשר לשוני שנחוץ לצורך קביעת תפקיד הפועל, ובדיקה בחלון גדול יותר נותנת הקשר רב יותר. עם זאת, דווקא לא מצאנו הבדל משמעותי מבחינת המאפיינים-נראה

שהתפקיד המורפולוגי והתפקיד התחבירי חשובים באותה מידה. בכל מקרה, גם לתנ"ך, גם לעברית המודרנית וגם למאגר המשולב המסווג הכי טוב נמצא עבור חלון של שתי מילים ועבור מלוא המאפיינים שהשתמשנו בהם.

קיים הבדל משמעותי מבחינת איכות הסיווג בין התנ"ך לעברית המודרנית-בעוד שבתנ"ך יש F Score של 0.32-0.35, בעברית המודרנית ה-F Score גדול יותר בכ-0.1, שזה מקביל לתוספת של 10% לאחוז הדיוק. אחוזי הדיוק אמנם נראים נמוכים ממבט ראשון, אך יש לציין שהם טובים פי כמה מניחוש מכיוון שלפי שיטת גלינרט יש 8 משמעויות שונות לבניין, כלומר ה-F Score שיהיה לניחוש רנדומלי הוא בעל תוחלת 0.125. חשוב לציין גם שבניין התפעל הוא הכי "קשה"-יש בו יותר משמעויות מאשר לכל בניין אחר בשפה, על פי שתי השיטות.

תובנה מעניינת שמצאנו בנוסף-כמעט תמיד, המסווגים הפשוטים יותר היו מדויקים מהמסווגים המורכבים. בקוד המוגש אנו משתמשים בשלושה סוגי מסווגים-SVM, השכן הקרוב ועץ החלטות. למעשה ניסינו להשתמש במסווגים נוספים כמו רשת נוירונים ובייס נאיבי אך החלטנו לא לכלול אותם בקוד עקב ביצועים ירודים. האם עבור 1000 דוגמאות בלבד המסווגים המסובכים נכנסים למצב של overfitting? ניתן יהיה לבדוק את האפשרות הזאת באמצעות בדיקת אותם מסווגים על מאגר דוגמאות גדול יותר והשוואה של התוצאות.