Hebrew Treebank 1.0

מאגר נתונים מנותח תחבירית - גרסה 1.0

רקע

המאגר הלשוני המנותח תחבירית כולל 4,783 משפטים (88,747 מילים, 120,213 מורפימות) של ידיעות מעיתון 'הארץ', עם הפרדה מלאה של מילים למורפימות וניתוח צורני-תחבירי מלא. תכוניות מורפולוגיות שאינן רלוונטיות באופן ישיר למבנה התחבירי, כגון שורשים, בניינים ומשקלים, אינן מנותחות. האתר כולל גם גירסא של המאגר התחבירי שכוללת את הרמה המורפולוגית בלבד.

מאגר העצים נבנה עם חלקי דיבר וקטגוריות תחביריות קרובים ככל שניתן לאלו הקיימים ב .English Penn Treebank-הבדל מרכזי מאנגלית הוא שבמאגר העצים בעברית המילים המנותחות מופרדות למורפימות, שלכל אחת מהן עשוי להיות חלק דיבר ומיקום תחבירי משלה. לדוגמא, שתי המילים בבית הגדול מנותחות כחמש מורפימות: ב ה בית ה גדול, כשהמופע הראשון של מורפימת היידוע ה אינו גלוי במילה בבית.

חלוקה זו למורפימות מאפשרת לנתח מורפימות שונות במילה כשייכות לרכיבים שונים בעץ התחבירי:

[ב [[ה בית] [ה גדול]].

הבדל מרכזי נוסף מאנגלית הוא הסדר החופשי יחסית של רכיבים בעברית. כדי לאפשר את קידוד התפקיד התחבירי של רכיב בתוך מבנה עץ נעשה שימוש בתכוניות פונקציונליות (נושא, מושא וכו') של רכיבים, כשהם עצמם מופיעים בתוך מבנה "שטוח."

מספר קטגוריות נוספו לאלו של ה Penn Treebank-על מנת לתאר מאפיינים מיוחדים של עברית כמו צורות הסמיכות, סימון יחסת המושא (את), מורפימת היידוע ה, ופרדיקטים נטולי פועל (במשפטים שמניים או תאריים.(

לתיאור מפורט של סכימת התיוג ראו סימעאן ואחרים. לדוגמאות ומוסכמות נוספות ראו את המדריך לתיוג.

תוכנה

לצורך פיתוח המאגר, נעשה שימוש בתוכנות הבאות:

- **SEMTAGS - Remko Bonnema, University of Amsterdam**סביבה גרפית לניתוח עצים
- **מנתח מורפולוגי (אראל סגל, טכניון)**מספק תיוג התחלתי למילות הקלט, אשר מתוקן ידנית ע"י המתייגים האנושיים
- **תוכנית להמרת פורמטים (אלון אלטמן, טכניון)**המרת הפלט של המנתח המורפולוגי לפורמט המתאים למאגר, ויצירת עצים התחלתיים שטוחים לניתוח

- **מיפוי עצים לטקסט המקורי (אביחי דגני, טכניון)**התוכנית ממפה כל מילה בטקסט המקורי למורפמות המתאימות לה בעץ הגזירה, ובודקת את ההתאמה בין הטקסט לעצי הניתוח

הערות

עצים ריקים: הטקסט המקורי חולק למשפטים באופן אוטומטי. בחלק מהמקרים משפט אחד בטקסט המקורי חולק למספר משפטים. לפיכך, כחלק מניתוח העצים היה צורך לאחד מחדש את אותם משפטים. כדי לשמור על הסנכרון בין מספרי המשפטים (שהתקבלו מהחלוקה האוטומטית) לבין מספרי העצים, הוכנסו עצים ריקים. לדוגמה אם המשפט הראשון פוצל לשלושה משפטים, 1,2 ו3, אז עץ מס' 1 יכיל את שלושת החלקים, ואילו עץ 2 ועץ 3 יהיו עצים ריקים.

עץ ריק נראה כך:

S
|
yyDOT
|
yyDOT

ומיוצג ע"י((S (yyDOT yyDOT)) :

משפטים כפולים: משפטים 24-36 בטקסט המקורי לא מופיעים במאגר, מאחר שהם חוזרים על עצמם במשפטים 1358-1370. משפטים 1249-1293 לא מופיעים במאגר, מאחר שהם חוזרים על עצמם במשפטים 1204-1248.

עצים חסרים: העצים הבאים לא מופיעים כרגע במאגר (במקומם מופיעים עצים ריקים): 552,772,1350,1382,2044,3206.

המרה בין תעתיק עברי ולטיני:

א	ב	ג	ד	ה	ו	ז	ח	ט	י	ך	כ	ל	ם	מ	ן	נ	ס	ע	ף	פ	ץ	צ	ק	ר	ש	ת	"	%
A	B	G	D	H	W	Z	X	J	I	K	K	L	M	M	N	N	S	E	P	P	C	C	Q	R	F	T	U	O

צוות

- **הנחיה** פרופ' אלון איתי, פרופ' יועד וינטר, ד"ר ח'ליל סימעאן
- **פיתוח הסכימה וניתוח** נעמי גוטמן, פנינה וייסברג, רוני טנצמן, עדי מילאה, נועה נתיב, יאיר עדיאל, שירי קינן
- **ליווי טכני** אלון אלטמן, רועי בר-חיים ,**שלמה יונה**