Hebrew Treebank 2.0

מאגר נתונים מנותח תחבירית - גרסה 2.0

רקע

מאגר הנתונים גרסה 2.0 כולל 6500 משפטים של ידיעות מעיתון 'הארץ', עם סימון מלא של חלוקת מילים וניתוח מורפולוגי-תחבירי. תכוניות מורפולוגיות שאינן רלוונטיות באופן ישיר לניתוח תחבירי, כגון שורשים, בניינים ומשקלים, אינן מנותחות.

ניתן גם להוריד גרסה הכוללת רק ניתוחים ברמת המילה, ללא הניתוח התחבירי.

גרסה 2.0 של המאגר המנותח כוללת שלושה שיפורים/תוספות ביחס לגרסה הקודמת:

תיקון והאחדה של ניתוחים קודמים.

תוספת של 1500 משפטים מנותחים.

ובמיוחד - ציון יחסי אב-בן (ראו בהמשך.)

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).

***חדש* - תלויות אב-בן**
גרסה 2.0 כוללת סימון תכוניות מיוחד לתיאור כל אותם מקרים שבהם התכוניות המורפו-תחביריות של צומת באות בירושה מאחד או יותר מבניו. אנו קוראים למצב כזה "תלות אב-בן". במבנה שבו Xהוא צומת ו Y-הוא אחד מבניו של X, תלות בין X ל Y-

מסומנת ע"י הוספת התכונית _DEP לצומת ,Y כאשר Z הוא סימן המציין את סוג התלות. אנו מבחינים בין ששה סוגים של תלויות:

DEP_HEAD: Y הוא הבן היחיד של X שקובע את תכוניות .X למשל :
זה היחס בין שם עצם פשוט ('בית') שאינו בצורת סמיכות ובין צירוף שמני בלתי מיודע ('בית גדול'). יש לשים לב שתלויות ראש כאלה אינן בהכרח מסומנות כאשר אין תכוניות המועברות מבן לאב.

DEP_DEFINITE: Y קובע אך ורק את תכונית הידוע של .X לדוגמה :
זהו היחס בין ה' הידוע על שם עצם פשוט וצירוף שמני שאינו בצורת הסמיכות הכולל אותו ('הבית הגדול'). דוגמה חשובה אחרת לתלות של ידוע היא היחס בין סומך בתוך צירוף סמיכות לבין כל צירוף הסמיכות. (ראו בהמשך.(

DEP_NUMBER: Y קובע אך ורק את תכונית המספר של .X לדוגמה :
זהו היחס בין מילת המספר 'אחד' והצירוף השמני המכיל אותה במבנה פרטיטיבי כמו 'אחד הבתים'. יחס זה מתאר את תלותו של מספר הצירוף השמני (יחיד) במילת המספר 'אחד', ולא בשם העצם ('בתים.('

DEP_MAJOR: Y קובע את רוב תכוניותיו של ,X אבל אחד או יותר מאחיו של Y קובע את הידוע ו/או את המספר של .X לדוגמה :
זהו היחס בין שם עצם נסמך ובין כל צירוף הסמיכות. זאת שכן (רק) הידוע של צירוף סמיכות נקבע ע"י הסומך, ולא ע"י הנסמך. מוסכמות אלה מובילות לניתוח הבא של הסמיכות 'בית הילדה:'

(NP-ZR-H

(NNT-ZY-DEP_MAJOR בית)

(NP-NY-H-DEP_DEFINITE

(H-DEP_DEFINITE ה)

(NN-NY ילדה)))

באופן דומה, במבנה פרטיטיבי:

(NP-ZY

(CDT-ZY-DEP_NUMBER אחד)

(NP-ZR-H-DEP_MAJOR

(H-DEP_DEFINITE ה)

(NN-NY בתים)))

DEP_ACCUSATIVE: Y קובע את תכונית ה-OBJ (מושא) של X. זהו היחס בין סימן היחסה 'את' והצירוף השמני הכולל אותו.

DEP_HEAD_MULTIPLE: Y הוא אחד מכמה בנים של X שקובעים את תכוניותיו במבנה מקבילי. זה המקרה ברוב ה-X-ים ששולטים על מבנה מחובר, כמו גם בכמה צירופי יחס מיוחדים.

תיעוד נוסף

לתיאור מפורט של סכימת התיוג ראו סימעאן ואחרים (2001):
http://mila.cs.technion.ac.il/treebank/tal.pdf

.לדוגמאות ומוסכמות נוספות ראו את המדריך לתיוג.

תוכנה

לצורך פיתוח המאגר, נעשה שימוש בתוכנות הבאות:

SEMTAGS - Remko Bonnema, University of Amsterdamסביבה 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:

		ת	ש	ר	ק	צ	ץ	פ	ף	ע	ס	נ	ן	מ	ם	ל	כ	ך	י	ט	ח	ז	ו	ה	ד	ג	ב	א
		t	e	r	q	c	c	p	p	y	s	n	n	m	m	l	k	k	i	v	x	z	w	h	d	c	b	a

צוות

הנחיה פרופ' אלון איתי, פרופ' יועד וינטר, ד"ר ח'ליל סימעאן

פיתוח הסכימה וניתוח ד"ר יובל קרימולובסקי, נעמי גוטמן, פנינה וייסברג, רוני טנצמן, עדי מילאה, נועה נתיב, יאיר עדיאל, שירי קינן

ליווי טכני ועיצוב ד"ר יובל קרימולובסקי, אלון אלטמן, רועי בר-חיים