**תאריך: 9 בינואר 2020**

**קורפוס השפה העברית – פרויקט מעבדה**

רשות התקשוב הממשלתי, החליטה ליזום פרויקט שעיקרו קידום השפה העברית על ידי ביצוע "דיגיטציה" של השפה, באופן שיאפשר שימוש בקורפוס שיתויג לצורך עיבוד שפה טבעית (NLP).

רשות התקשוב הממשלתי התקשרה עם האקדמיה ללשון העברית, שהיא [המוסד העליון ללשון העברית](http://hebrew-academy.org.il/wp-content/uploads/acad-law.pdf) על פי חוק ובעלת ניסיון של כשישים שנה בניתוח מורפולוגי של קורפוסים ממוחשבים, לצורך ביצוע פרויקט מקדים במתכונת של מיזם מעבדה. מטרת פרויקט המעבדה הייתה לבחון את האתגרים בתיוג ידני של קורפוס בשפה עברית בת-זמננו, לפי תקן UD. תקן זה הוא תקן אוניברסלי, שמאפשר שימוש על ידי אלגוריתמים שונים לעיבוד שפה טבעית.

על מנת להבטיח קישור בין עולם התיוג המורפולוגי לבין עולם הניתוח התחבירי הממוחשב, שולבה בפרויקט [מעבדת ה- NLP של פרופ' רעות צרפתי](http://nlp.biu.ac.il/~rtsarfaty/onlp/hebrew/) באוניברסיטת בר-אילן.

לצורך מיזם המעבדה נבחר קורפוס שכולל כ- 600 משפטים בעברית, מעולם התוכן של **משרדי הממשלה** (דפי תוכן באתרי הממשלה) **והשירות לציבור** (פניות בדואר אלקטרוני למוקד השירות הממשלתי, לאחר מחיקת פרטים מזהים).

תוצרי הפרויקט הוגדרו מלכתחילה ככאלו שיפורסמו באתר Data-Gov ויועמדו לרשות הציבור הרחב, חוקרי השפה העברית וחברות מסחריות.

האקדמיה ללשון העברית תייגה ידנית את המשפטים האלה בתיוג מורפולוגי על פי תקן UD.

להלן פירוט העמודות שבקובץ התיוג המורפולוגי:

**ID (word index) -**

* **sent\_id:** מספר סידורי של המשפט, המספר לפני הנקודה מתאים לקובץ המטא-נתונים;
* **text:** המשפט המלא;
* **מספר:** מספר סידורי של המילה במשפט.

**U-POS** – חלק הדיבר של המילה במשפט לפי תקן UD.

**X-POS** – חלק הדיבר בייצוג ישן (זהה לעמודת ה-U-POS).

**FORM** – צורת המילה כפי שהיא במשפט המקורי.

**LEMMA** – הערך המתאים על פי רשימת הערכים בתקן ה-UD הקיים של העברית.

**FEATS** – התכונות המורפולוגיות של המילה.

**MISCELLANEOUS** – שונות.

**LEMMA ACADEMY** – הערך המתאים והמנוקד של המילה על פי האקדמיה ללשון העברית.

המעבדה לעיבוד שפה טבעית של פרופ' רעות צרפתי באוניברסיטת בר-אילן השתלבה במיזם המעבדה הנוכחי כדי להניח את הרובד התחבירי על יסוד הניתוח המורפולוגי הידני.

קובץ התיוג התחבירי המובא כאן מתויג באופן אוטומטי בלבד (לאחר הזנה של התיוג המורפולוגי הידני) ולא נבדק לאחר מכן באופן ידני. הקובץ הוא תוצר של מערכת ה-YAP ועל כן מתויג בתקן SPMRL (Statistical Parsing of Morphologically Rich Languages). שימו לב כי קובץ התיוג התחבירי אינו מכיל את מספרי המשפטים (ID), ועל כן אין אפשרות לקשר בין הקובץ הזה לבין קובץ המטא-נתונים של המשפטים. כמו כן, קובץ התיוג התחבירי אינו כולל את משפט המקור.

להלן פירוט העמודות שבקובץ התיוג התחבירי:

**ID** – מספר סידורי של המילה במשפט.

**FORM** – צורת המילה כפי שהיא במשפט המקור לאחר פירוקהּ (Parsing).

**LEMMA** – הערך המתאים על פי רשימת הערכים של מערכת ה-YAP.

**XPOS** – חלק הדיבר לפי ייצוג ישן.

**UPOS** – חלק הדיבר לפי תקן SPMRL.

**FEATS** – התכונות המורפולוגיות של המילה.

**HEAD** – האינדקס (ID) של המילה שממנה יוצאת קשת תחבירית למילה הנוכחית.

**DEPREL –** שם הקשת התחבירית.