האקדמיה ללשון העברית

טיוטה לתקן האקדמיה ללשון העברית מספר ..... Draft IHA.....

2022

טיוטה לתקן האקדמיה ללשון העברית – התאמת תקן UD לשפה העברית

מסמך זה הוא התאמת תקן UD הבין־לאומי לצורכי השפה העברית.

מסמך זה הוא מסמך טיוטה להערות בלבד ואינו תקן מאושר של האקדמיה ללשון העברית.

האקדמיה ללשון העברית

את תקן האקדמיה הזה (להלן – התקן[1]) הכינה ועדת התקן שמינתה האקדמיה ללשון העברית, בהרכב זה:

נועם אורדן; יראל אושרת; אורלי אלבק; אבנר אלגום; רועי אלמוג; יפעת בן-משה; ענת בר סימן טוב; יעקב גוטקין (משקיף); איילת הראל (יו"ר); נתנאל דהן; עופר ישי; יעל נצר; יובל פינטר (משקיף) רעות צרפתי; נריה רבלין; דורון רובינשטיין (משקיף); אבי שמידמן

הקשר בין התקן למסמכים בין־לאומיים: התקן מתבסס על תקן UD[2] ופורטו בו ההתאמות לשפה העברית.

עדכניות התקן: התקן יעודכן על פי הצורך. המשתמשים יוודאו שבידיהם המהדורה המעודכנת של התקן.

תוקף התקן: התקן נכנס לתוקף במועד פרסומו באתר האקדמיה ללשון העברית. גרסה באנגלית של התקן על פי התצורה המקובלת של UD תתפרסם באתר UD אחרי פרסום התקן בעברית.

זכויות היוצרים:

- **זכויות היוצרים בתקן הן של האקדמיה ללשון העברית;**
- **התקן הוא תקן פתוח – מותר** לצלם, להעתיק או לפרסם בכל אמצעי שהוא את התקן או קטעים ממנו, ובלבד שיצוין כי זכויות היוצרים הן של האקדמיה ללשון העברית.

[1] השימוש במונח "תקן" נעשה לצורכי נוחות. יובהר כי מדובר בתקן של האקדמיה ללשון העברית ולא בתקן על פי חוק התקנים, תשי"ג–1953, או תקן ישראלי של מכון התקנים הישראלי.

[2] Universal Dependencies

האקדמיה ללשון העברית

1. מבוא ומטרת התקן

Universal Dependencies, הידוע בקיצור UD, הוא מיזם בין־לאומי משותף ליצירת עצי גזירה (treebank) לשפות העולם. המטרה העיקרית של המיזם היא להגיע לתיוג תואם בשפות רבות, אך גם לאפשר הרחבות ייחודיות לכל שפה.

מסמך זה בא להסביר כיצד נעשָׂה התיוג בשפה העברית, להשלים את החסרים שיש בתקן הבין־לאומי בקשר לעברית ולהוסיף את העניינים הייחודיים של השפה.

2. תהליך העבודה להכנת התקן

מסמך זה מתבסס על מסמך התיעוד של תקן UD בעברית שנמצא כאן.

כל סטייה מן המסמך המקורי מצוינת בהערה.

המסמך הוגש לוועדה להסדרת תקן UD בעברית בראשות האקדמיה ללשון העברית ובהשתתפות נציגי מערך הדיגיטל הלאומי - רשות התקשוב הממשלתי, חוקרי עיבוד שפה טבעית ונציגי האיגוד הישראלי לטכנולוגיות שפת אנוש כמפורט לעיל.

הוועדה דנה בעניינים המפורטים במסמך והכריעה בנושאים שונים.

לאחר אישור המסמך השלם בוועדה יפורסם התקן להערות הציבור.

3. הגדרות

ככלל התקן מסתמך על ההגדרות הכלליות לחלקי הדיבר ולתכוניות השונות המוגדרות בתקן הבין־לאומי.

4. מבנה המסמך

במסמך שני חלקים:

4.1 הסבר כללי על חלוקת המילים ליחידות בסיס לתיוג בשל מורכבותה המורפולוגית של העברית;

4.2 שינויים ותוספות על התקן הבין־לאומי בחלק המורפולוגיה של העברית.

5. חלות התקן

5.1 התקן מתפרסם במסגרת מיזם הקמת קורפוס השפה העברית בת־זמננו של מערך הדיגיטל הלאומי - רשות התקשוב הממשלתי ושל האקדמיה ללשון העברית.

5.2 התקן מגדיר את הדרישות האחידות לתיוג במיזם זה ובכל מיזם אחר מטעם ממשלת ישראל שיבוצע ישירות או באמצעות גורמים מתוקצבים.

5.3 אימוץ התקן הוא תנאי להשתתפות במיזמים ממשלתיים לתיוג ומומלץ גם בעבור גופים מחוץ לממשלה.

6. התקן בעברית – חלק ראשון: הקטעה (סגמנטציה) וחלוקה ליחידות בסיס (טוקניזציה)

6.1 ככלל מילה בעברית מתוחמת ברווחים.

6.2 סימני הפיסוק מיוצגים כתֵיבות (״תמנית״) נפרדות.

6.3 מילה שהיא צירוף מוקף (כגון **בין־משרדי**) תופרד לשלוש תיבות. קיום או אי-קיום הרווחים בין המילים האלה יבוא לידי ביטוי בתכונית SpaceAfter=Yes/No.

6.4 מילה שהיא **ראשי תיבות** לא תפורק לרכיביה אלא תובא כתֵיבה אחת ותקבל ערך אחד ובתכוניות יצוין Abbr=Yes.

6.5 אותיות **מש״ה וכל״ב** יחולצו מהמילה בטקסט ויובאו כתֵיבות נפרדות. אפשר שיבואו כמה מאותיות השימוש ברצף, ויש להפרידן אפילו אינן נראות לעין (לדוגמה: **בַּבַּיִת** יתפרק לרכיבים ב+_ה_+בית). יוצא מן הכלל צירוף האותיות **כש** שיכול להיות מופרד לרכיביו (כשהכוונה היא ״כפי ש״) או יכול לבוא בלי הפרדה כמילת שעבוד.

כדי לציין שהאותיות הללו מצטרפות בטקסט למילת הבסיס, מציגים את הרכיבים המצטרפים בעמודת הצורה ולפניהם, אחריהם או משני צידיהם נותנים קו תחתון צמוד בכיוון ההצטרפות.

6.6 **כינויים חבורים** יחולצו מהמילה בטקסט והמילה תפורק לתֵיבות נפרדות המייצגות את כלל הרכיבים בתוכה. הכינויים ומילות היחס יפורקו לערכיהם, ואילו המילה שאליה הצטרפו תישאר בצורתה בטקסט (לדוגמה: **שלנו** יתפרק לרכיבים של_ + _אנחנו; **ביתו** יתפרק לרכיבים ה_+_בית_ +_ של_ + _הוא; **אהבתיה** יתפרק לרכיבים אהבתי_ + _את_ + _היא). כדי לציין שהכינויים ומילות היחס הללו מצטרפים בהיתוך מורפולוגי בטקסט הגולמי, מציגים את הרכיבים המותכים בתוספת קו תחתון כפי המוסבר לעיל בסעיף 6.5.

6.7 בצורת **שם המספר** + **כינוי גוף** (דוגמת **שלושתם**): יש לפרק את המילה וכל רכיב יקבל את חלק הדיבר שלו: שלושה_ + _הם. אין להוסיף "של" בפירוק.

6.8 **תואר הפועל** + **כינוי גוף** (דוגמת **לבדם**, **לאיטו**): יש להפריד את כינוי הגוף ולתת לכל רכיב את חלק הדיבר שלו: לבד_ + _הם (ADV, PRON בהתאמה).

6.9 **כינויי השאלה** + **כינויי גוף** (דוגמת **מיהם**, **איזהו**): יש להפריד את המילה לרכיביה ולתת לכל רכיב את חלק הדיבר שלו.

6.10 **הכמתים** + **כינויי גוף** (רובו, כולו, מקצתו, מרביתו): יש להפריד את המילה לרכיביה ולתת לכל רכיב את חלק הדיבר שלו (למשל: כל_ + _הוא), מבלי לציין יחסה על כינוי הגוף בתכוניות. הקשת התחבירית תישאר כפי שהייתה כיוון שעניין פירוק הכמתים אינו מפריע לצד התחבירי.

6.11 צורות דוגמת **סבורני**: יש להפריד את המילה לרכיביה (סבור_ + _אני) בלא ציון יחסה על כינוי הגוף.

6.12 **הינה** + **כינוי גוף** (דוגמת **היננו**, **הינך**, **הינו**): יש להפריד את המילה לרכיביה תמיד.

6.13 **יש ואין** + **כינויי גוף** (דוגמת: **ישנו**, **ישנם**, **אינם**): אין לפרק את המילה. כינוי הגוף יתווסף כתכונית.

האקדמיה ללשון העברית

7. התקן בעברית – חלק שני: מורפולוגיה – חלקי הדיבר (U-POS) והתכוניות (Feats)

7.1 החלטות שהתקבלו בנושאים שונים:

7.1.1 מילה שהיא ראשי תיבות (מקבלת בתכוניות Abbr=Yes) – במקרה שאי אפשר לקבוע חלק דיבר אחד לראשי התיבות, יינתן חלק דיבר X (למשל בראשי התיבות זש"ה [=זהו שאמר הכתוב]).

7.1.2 **יש** תקבל חלק דיבר VERB בתוספת התכונית Existential=Yes. הלֶמה של המילה תהיה 'יש'. **אין** הקיומית תקבל גם כן חלק דיבר VERB.

7.1.3 **אין לא קיומית (=אין שוללת)**: בצורה הנוטה (**אינו**, **אינך** וכו') – תמיד תקבל AUX; בצורה הגרודה (**אין**) – תמיד תקבל ADV. הלֶמה בשני המקרים תהיה 'אין'.

7.1.4 צורות **בינוני פעול** יקבלו תמיד את חלק הדיבר ADJ. לדוגמה: היום **רצוי** שכולם יסכימו, היום **רצוי** להסכים.

7.1.5 צורות עם גוון מודלי/אספקטואלי כרכיב ראשון ב"נשוא מורחב" יתויגו VERB או ADJ לפי התנהגותן בשפה: אם אפשר להטות את הצורה בזמנים היא תתויג VERB, אם לא – ADJ.

לדוגמה: דני **מוכן** לעשות את זה מחר (≠ דני **הוכן** לעשות את זה מחר) >> ADJ;

דני **עומד** להגיע בכל רגע (= דני **עמד** להגיע בכל רגע) >> VERB;

דני **חייב/צריך/אמור** לעשות את זה (דני **היה** חייב/צריך/אמור לעשות את זה) >> ADJ.

7.1.6 תיוג הערך **היה**

א. **היה** שמתחלף בקלות עם "יש" תתויג VERB.

לדוגמה: פעם **היו** דינוזאורים (= היום **יש** דינוזאורים) >> VERB

לישראל **היו** שלושה נציגים (= לישראל **יש** שלושה נציגים) >> VERB

ב. **היה** שאין לו משלים ויש לו משמעות קיומית ברורה יתויג תמיד VERB.

לדוגמה: זה לא **היה** >> VERB

זה יכול **להיות** >> VERB

ג. כל **היה** אחר יתויג AUX כפי שמפורט להלן:

- **היה** הבא לפני פועל אחר בצורת בינוני לציון אספקט (פעולה מתמשכת/הרגלית) או מודוס (איווי או מצב קונטרה-פקטואלי).

  לדוגמה: בעבר **היינו** הולכים לשם הרבה >> AUX

  הייתי מאמצת חתול עכשיו >> AUX

- **היה** שאינו מתחלף ב"יש" ומצוי במשפטים עם איבר היכול להיתפס כנשוא ש"ע/ש"ת.

  לדוגמה: הסכין **היה** חד >> AUX

  הוא **היה** מורה להתעמלות >> AUX

  הוא רוצה **להיות** דינוזאור (= הוא רוצה שהוא **יהיה** דינוזאור) >> AUX

- **היה** שאינו מתחלף ב"יש" ומצוי במשפטים עם איבר היכול להיתפס כנשוא אדוורביאלי.

לדוגמה: הסכין **היה** במטבח >> AUX

זה **היה** בטעות >> AUX

המסיבה **הייתה** אתמול >> AUX

7.2 נושאים נוספים שטרם סוכמו ושיועלו לדיון בוועדה בהמשך:

7.2.1 תיוג כינויי הרמז (PRON או DET);

7.2.2 תיוג החג"מים (כדאי, אפשר) והמילים בעלות גוון חג"מי;

7.2.3 תיוג צורות הבינוני (מתי שם עצם, מתי שם תואר ומתי פועל ממש?);

7.2.4 סימון התכוניות השונות (תכונית היחסה, היידוע, המין הדקדוקי, המספר, הקיום וכו');

7.2.5 תיוג תוארי פועל שהם צירוף.

8. תחביר

טיוטת ההצעה לוועדה אינה כוללת את פרק התחביר.

9. בנק עצים

להלן בנק העצים הקיים בעברית:

https://universaldependencies.org/treebanks/he_htb/index.html