UCCA: Hebrew annotation and tokenization guidelines

1. Annotation guidelines

1. Double genitive construction:

    - Beit_C o_F [shel_S Dani_A (Beit)_A]_E (בית-ו של דני)
    - [Dmut__C o_F]_E [shel_R nachash_C]_C (דמות-ו של נחש)
    - Sovlanut_S o_F [shel_R Dani_C]_A azla_D (סובלנות-ו של דני אזלה)

2. Static scenes

    i. [zo_A [ha_F machberet_C [shel_S Yael_A (machberet)_A]_E]_A (IMP)_S]_H (זו ה-מחברת של יעל)
    Zero copula

    ii. Haita_S zo_A [tmunat nachash]_A (הייתה זו תמונת נחש)

    iii. Ma_A ze_S [ha_F davar_C [ha_F ze_C]_E]_A? (מה זה ה-דבר ה-זה)

    iv. ‘Et_A ze_S [kli ktiva]_A (עט זה כלי כתיבה)

    v. [Ein_D/S zo_A kivsa_A]_H (אין זו כבשה)

    vi. Haita_S [l_R o_C]_A kivsa_A (הייתה ל-ו כבשה)

3. Imperatives

Imperatives in Hebrew do not require adding an IMP unit; the imperative itself can be marked as the A since it is conjugated for person.

- [Shlach]_P/A et ha michtav ba doar (שלח את ה-מכתב ב-דואר)
- [Shilchi]_P/A et ha michtav ba doar (שלחי את ה-מכתב ב-דואר)

4. PP Adverbials
We internally annotate PP adverbials as R+C:

- be_R hechlet_C (ב-החלט)
- me_R olam_C (מ-עולם)

- me_R chadash_C (מ-חדש)
- ka_R rauy_C (כ-ראוי)

In rare cases where an expression is completely opaque, it should be left unsegmented at the tokenization stage. If at the annotation stage you encounter a word that was wrongly segmented, you can use the re-tokenize feature to contract it back

5. Reflexive ‘atzmi’ (עצמי):

    - Dani_A [rachatz ‘atzm o UNA]_P (דני רחץ עצמ-ו)
    - [ani_C [‘atzm i UNA]_F]_A lo_D haiti_F ‘ose_P [davar_C [ka ze]_E]_A (אני עצמ-י לא הייתי עושה דבר כ-זה).

6. Levadi (לבדי) is marked D, internally it is unanalyzable:

    - Dani_A ‘asa [levad o UNA]_D et ha mesima (דני עשה לבד-ו את ה-משימה)

7. Oto (אותו)

    - As a direct object it is internally analyzable: Pagashti_P/A [ot_R o_C]_A (פגשתי אות-ו)
    - When conveying sameness it is internally unanalyzable: [Be_R [ot o UNA]_E ha_F sefer_C] (ב-אות-ו ה-ספר)

8. Annotation of Participants:

    In Hebrew, a Participant may appear as:

a. A free word:
    - Dani_A halach_P [la_R gan]_A (דני הלך ל-גן)
    - Hu_A nimtza sham (הוא נמצא שם)

b. A verb conjugated for person:
    - Rainu_P/A seret_A (ראינו סרט)
    - Shamati_P/A kol_A [ba_R chutz_C]_A (שמעתי קול ב-חוץ)

c. Pronominal suffix:
    - Savlanut_S i_A (סבלנות-י)
    - Hadavar_A [lakad et ‘eina]_P v_A (ה-דבר לכד את עיני-ו)

In some cases, a single scene may contain multiple referents to the same Participant. We will then need to determine which referent to select as the A :

a. Prefer marking a free word as the A over conjugated verbs and suffixes; any other referent which was not marked as the A should then be marked F.

- Hu_A chazar_D [‘al bakashat]_P o_F (הוא חזר על בקשת-ו)
- Hu_A [taman et yad]_{P-} o_F [ba tzalachat]_{-P} (הוא טמן את יד-ו ב-צלחת)

b. If there is no free word that can serve as the A, we prefer marking a conjugated verb as the A and then any other referent will be marked F.

- Nisiti_D/A [lehotzi mitachat yad]_P i_F tziur_A (ניסיתי להוציא מתחת יד-י ציור)

c. If the only referent of a Participant is a pronominal suffix then it should be marked as the A:
[Chaschu ‘eina]_P v_A (חשכו עיני-ו)

9. Remotes:

Generally, we prefer to add free words as Remotes over pronominal suffixes, but if there isn’t a free word that can be added, it is OK to add a pronominal suffix.

- [Hora_S/A v_A]_A amru_P [l_R o_C]_A [lehafisk_D le’ashen_P (o)_A]_A (הורי-ו אמרו ל-ו להפסיק לעשן)

10. Purposive linkage:

Sometimes the infinitive “ל” can express a purposive linkage, but since we do not segment it, it cannot be marked as the Linker. In such cases add an IMP L instead.
[Dani_A halach_P [le_R merkaz_C ha ’ir_C]_A]_H (IMP L) [lirot_P [et_R ha_F hatzaga_P]_A (Dani)_A]_H (דני הלך ל-מרכז ה-עיר לראות את ה-הצגה)

11. Negative polarity items

We will prefer marking the negative word “לא” as the D over any negative polarity item that may also occur in the same scene (e.g. אף אחד, שום דבר, כלום, מאומה)

- Lo_D raiti_P/A [klum]_A

12. Ein/lo+ela construction:

a) when it is used as a contrasting negation construction (evokes two scenes) we mark ein/lo as D and ela as L:

i) [ein_D [ha yeladim]_A holchim_P [la migrash]_A]_H ela_L [[le beit sefer]_A (holchim)_P (yeladim)_A] (אין ה-ילדים הולכים ל-מגרש אלא ל-בית -ספר)

ii) [Dani lo_D ochel_P 'agvania_A]_H ela_L [melafefon_A (Dani)_A ochel_P]_H (דני לא אוכל עגבניה אלא מלפפון)

b) when it is used as an emphasis construction that does not evoke two separate scenes, we mark "ein/lo..ela" as a discontiguous UNA D:

- [ein]_{D-} [l o_]_A [ela]_{-D} [dvarim tovim]_A lehagid_P (אין ל-ו אלא דברים טובים להגיד)

13. Interesting examples

a) Different uses of body parts:

- [ha_F ma'ase]_A orer_D [be_R lib_C o_C]_A regashot simcha_P (ה-מעשה עורר ב-לב-ו רגשות שמחה)
- [Ha_F yeled_C]_A [amar be lib]_P o_F (ה-ילד אמר ב-לב-ו)
- Dani [hita ozen]_P [le divrei ha more]_A (דני הטה אוזן ל-דברי ה-מורה)
- Dani_A nianea_P [rosh_C o_E]_A (דני נענע ראש-ו)

2. Tokenization guidelines

1) The following prefixes are segmented[1]:

ב,כ,ל,מ	**Prepositions**	דני הלך **ל** גן דני נמצא **ב** ביה"ס דני הלך **מ** שם יפה **כ** פרח Including where the prefix begins a prepositional phrase (PP) adverbial[2]: תנהג **ב** זהירות
ה	**Definite article**	**ה** ילד הלך מהר. Including when attached to demonstratives[3]:

[1] Note that we do not segment the infinitive "ל":
לראות, לשמוע

[2] Generally, prefer to segment prepositional prefixes wherever possible:
דני התחיל מ חדש, דני ביקש ש יתקשרו אל יו ב הקדם, דני ב החלט מעונין להשתתף , דני נמצא ב פנים
דני התנהג כ ראוי

[3] An exception would be הללו which should not be segmented.

המחשב ה זה עובד האיש ה הוא מוכר		
ה רואה אתה את מה ש אני רואה?	**Interrogative**	**ה**
הכלב ש דני ראה הוא שחור הילד ה יושב בשורה הראשונה הוא דני	**Relativizers**	**ש, ה**
דני ראה ש הכלב מתקרב. כש דני יבוא, נתחיל.	**Subordinating conjunctions**	**ש, כש, מש, לכש**
דני **ו** דנה	**Coordinating conjunction**	**ו**
דני מכר **כ** אלף פרחים	**Adverb**	**כ**

2) Pronominal suffixes are segmented:

a) *Possessive pronominal suffixes appended to nouns:*

ספר: ספר י, ספר ך, ספר ו, ספר ה, ספר נו, ספר כם, ספר כן, ספר ם, ספר ן
ספרים: ספר י, ספרי ך, ספרי ו, ספרי ה, ספרי נו, ספרי כם, ספרי כן, ספרי הם, ספרי הן
שפה: שפת י, שפת ך, שפת ו, שפת ה, שפת נו, שפת כם, שפת כן, שפת ם, שפת ן
שפות: שפות י, שפות יך, שפות יו, שפות יה, שפות ינו, שפות יכם, שפות יכן, שפות ם, שפות ן

b) *Pronominal suffixes as the objects of prepositions:*

בשביל: בשביל י, בשבילְ ך, בשביל ו, בשביל ה, בשביל נו, בשביל כם, בשביל כן, בשביל ם, בשביל ן
לפני: לפנ י, לפני ך, לפני ו, לפני ה, לפני נו, לפני כם, לפני כן, לפני הם, לפני הן
ל-: ל י, ל ך, ל ו, ל ה, ל נו, ל כם, ל כן, ל הם, ל הן
מן: ממ ני, ממ ך, ממ נו, ממ נה, ממ נו (מאת נו), מ כם, מ כן, מ הם, מ הן
את: אות י, אות ך, אות ו, אות ה, אות נו, את כם, את כן, אות ם, אות ן

c) *Pronominal suffixes as the direct objects of verbs*

הקיף הקיפ ני, הקיפ ך, הקיפ ו, הקיפ ה, הקיפ נו, הקיפ כם, הקיפ כן, הקיפ ם, הקיפ ן

d) *Pronominal suffixes appended to the infinitive construct:*

i) *as the subject of the infinitive:*

ב ראות י זאת, התחזקה ב י ה הרגשה

ii) as the object of the infinitive:

דני מנסה לראות ה כל יום

3) We segment contracted existential particles from their pronouns

אינ ני, אינ ך, אינ ו, אינ ה, אינ נו, אינ כם, אינ כן, אינ ם, אינ ן
יש נו, יש נה, יש נם, יש נן

4) We segment contracted reflexives "עצמי" and "לבדי":

עצמ י, עצמ ך, עצמ ו, עצמ ה, עצמ נו, עצמ כם, עצמ כן, עצמ ם, עצמ ן
לבד י, לבד ך, לבד ו, לבד ה, לבד נו, לבד ם, לבד ן

5) We segment the following blended forms: איזה, מיהו, מהו, זוהי, זהו,

a) זה ו ה בית של דני
b) זו הי ה תשובה ה נכונה
c) מה ו ה רעיון ה עומד מאחורי ה יוזמה?
d) מי הו האיש העומד שם?
e) איזה ו עשיר ה שמח ב חלק ו

6) Also, note that when conveying obligation, "על" in its contracted forms should be segmented as well:

על יו להשתתף
על ינו לעשות את הדבר הנכון

7) We do not segment the construct state (e.g. ספר**י** ילדים, שמל**ת** ה נשף)

8) We do not segment the binyan or subject agreement prefixes/suffixes of verbs (**הת**רחצ**תי**, **ת**למדו)

9) We do not segment gender/number suffixes on nouns and adjectives: ספר**ים** טוב**ים**, תלמיד**ה** טוב**ה**

10) We do not segment acronyms (e.g. צה"ל).

11) Note that because of an automatic tokenization process that occurs before you receive the passage, you may encounter a word that was unnecessarily tokenized for containing a geresh (e.g. ג'ונגל may appear ג ' ונגל). In such cases please delete the unnecessary space (should be simply ג'ונגל).

12) You might also see that due to automatic tokenization spaces occur between a makaf and the words it connects (בית - ספר). Those spaces are correct, so please do not delete them.

13) We segment prepositions in fixed expressions like ב בקשה.

14) We do not segment derivational morphemes like עיראקי, מהירות

15) We do not segment morphemes in borrowed words like אנטישמי.

16) We do not segment complex and compound prepositions (that are combinations of a preposition+noun or preposition+preposition). For example, the preposition "באמצעות" should not be segmented.

17) We do not segment compound conjunctions like מאחר ש-, מפני ש-

18) We do not segment compound questions words such as איפה, כמה, לאן and למה (note that both לָמָה and לְמָה should not be segmented)

Prepositions in the Hebrew Treebank (UD) and their Frequency

Use this list to identify prepositions that you are uncertain of. If you find in the text a preposition that does not appear on this list, please mention this in the comment field in the follow-up spreadsheet.

ב 7928
של 4830
ל 4438
את 1998
מ 1697
על 1503
כ 573
עם 457
בין 281
מן 213
עד 211
כדי 185
לפני 153
לאחר 148
אל 147
כמו 141
נגד 124
אחרי 114
ללא 71
לפי 56
לעומת 50
אחר 47
באמצעות 46
למרות 44
בגלל 44
בידי 43
מפני 39

38 בעקבות
38 במשך
37 תוך
36 בלי
35 ליד
35 לגבי
32 במקום
30 לבין
29 בפני
29 אצל
28 מול
28 בשל
27 מתוך
27 בתוך
23 לקראת
23 לעבר
22 למען
22 כלפי
22 בעוד
21 תחת
21 מעל
20 תמורת
20 בקרב
19 עבור
18 בשביל
18 בגין
17 סביב
16 מאחורי
16 בעד
15 מצד
15 בלא
14 מבין
12 לנוכח
11 מתחת
11 מעבר
11 לאורך
11 בטרם
11 בזכות
10 לשם
10 לידי
10 החל
9 מבלי
9 מאת
9 לצד
9 כש
9 כעבור
8 נוכח
8 כגון

7 הודות
6 ע"י
6 מלבד
6 כנגד
6 דרך
5 מבעד
5 לאור
5 באוזני
4 לזכות
4 חוץ מ
4 בעבור
4 בגדר
3 מש
3 לרגל
3 לפנות
3 לכבוד
3 בהתחשב
3 בתור
3 בצד
3 במו
3 בלעדי
2 פרט ל
2 סמוך ל
2 נוסף ל
2 משך
2 מחוצה ל
2 לתוך
2 למן
2 לכדי
2 כאל
2 בתוككי
2 בלית
1 קודם ל
1 משל (כמו "כאילו")
1 לצורך
1 למעין
1 לכעין
1 לדידי
1 כעין
1 זולת
1 במסגרת
1 בגנות
1 אודות

ביטויי יחס נוספים: על מנת, מחוץ, בפי