Ver. 3

הנחיות תיוג שמות ישויות בעברית

Hebrew Named Entity Annotation Guidelines

המשימה

המשימה היא לתייג אזכורים של *שמות ישויות* (Named Entity Mentions), המצביעים על ישויות השייכות לקטגוריות: אנשים (PER), ארגונים (ORG), ישויות גיאו-פוליטיות (GPE), מיקומים (LOC), מתקנים (FAC), מוצרים (DUC), יצירות (WOA), אירועים (EVE) ושפות (ANG).

שם-ישות הוא ביטוי שמצביע באופן *עצמאי* על *ישות ייחודית*. במשימה זו מתייגים שמות-ישויות השייכים לקטגוריות המוגדרות בלבד.

הביטויים המתוייגים יכולים להיות שמות ממשיים (proper names), ביטויים מיידעים (definite descriptions), ראשי תיבות, כינויים, קיצורים.

הביטויים המתוייגים יכולים להיות מקוננים. כששם ישות מכיל שמות ישויות נוספים, יש לתייג גם את הישויות הפנימיות.

הקטגוריות

אנשים (PER), ארגונים (ORG), ישויות גיאו-פוליטיות (GPE), מיקומים (LOC), מתקנים (FAC), מוצרים (DUC), יצירות (WOA), אירועים (EVE) ושפות (ANG).

a. **PER** - אנשים (Person) אינדיבידואליים. קבוצות של אנשים (כולל משפחות) הם לא ישות איש/ה. כולל דמויות בדיוניות (יש להבחין בין שם היצירה שתתוייג WOA לבין הדמות, למשל "בטמן" הסרט והדמות שמופיעה בו)

לרגל צאת סרטו החדש, [גבי עמרני] מנתץ מיתוסים...

לפי סקר שנערך לאחרונה, [צ'נדלר] הוא הדמות האהובה בסדרה...

הזוג נתניהו יעיד שוב במשפט דיבה.. (**אין תיוג** לקבוצות של אנשים)

b. **ORG** - ארגונים (Organization), סוכנויות, מוסדות וכל קבוצת אנשים המוגדרת בעזרת מבנה ארגוני מבוסס (חברות, עמותות, קבוצות ספורט, משרדי וגופי ממשלה, גופי בטחון, הרכבים מוזיקליים, גופי תקשורת וכו').

מניית [פייסבוק] צנחה אתמול...

במשחק הכדורגל בשבת [סכנין] ניצחה את [בני יהודה]...

חילופי קללות ב [וועדת הכספים] של [הכנסת]...

לעתים תהיה התייחסות לארגון על ידי שם חבריו ברבים, יש לתייג ORG - הדמוקרטים עבור המפלגה הדמוקרטית, הנאצים עבור המפלגה הנאצית. יש לשים לב שלא מדובר בהתייחסות כללית לקבוצת אנשים מסויימת, אלא כבינוי לארגון!

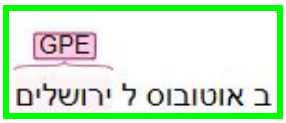
c. **GPE** - ישויות גיאו-פוליטיות (Geo-Political Entities) הן ישויות המורכבות מממשלה או שלטון, מיקום פיזי ואוכלוסיה. הסוגים הנפוצים הם מדינות, מחוזות, ערים, מועצות מקומיות, ישובים,

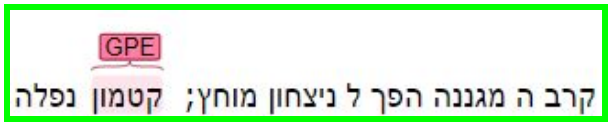
קיבוצים וכו'.
בתתי אזורים של GPE כמו "דרום [תל אביב]", יתוייג רק ה-GPE (ראה סעיף 5b להנחיות גבולות).

עשרות אנשים מגיעים מ[תאילנד] ל[ישראל]

[ארה"ב] תקפה את [עיראק]

הבהרה: אלה ישויות מורכבות בעלות שלטון, אוכלוסיה ושטח גיאוגרפי. הדוגמאות הקלאסיות הן מדינות וערים. חובה שכל השלושה יהיו קיימים כדי שהישות תיקרא GPE. באזכור של שם-ישות GPE, היא תתויג GPE בכל המקרים, בלי התחשבות באם הכוונה היא למיקום הגיאוגרפי, לאוכלוסיה או לשלטון. במשימה זו, שמות של שכונות - מתייגים כ-GPE.





לעתים תהיה התייחסות למדינה על ידי שם האוכלוסיה שלה ברבים, יש לתייג GPE - *"חרם דגנים על **[הסובייטים]**"*. יש לשים לב שלא מדובר בהתייחסות כללית לקבוצת אנשים, אלא כבינוי למדינה!

d. **LOC** - מיקומים (Location) הם אזורים המוגדרים בצורה גיאוגרפית או אסטרונומית, ***שאין להם רכיב פוליטי מוגדר***, או מבנים טבעיים כמו מקווי-מים או הרים. דוגמאות: גופים שמימיים, יבשות, אזורים וחבלי ארץ שלא מבוססים פוליטית ("הבלקן", "הנגב"), אוקיאנוסים, ימים, אגמים, מיצרים, איים, אגמים, שמורות טבע, הרים ורכסים.

רעידת אדמה הורגשה באזור [הכנרת]...

e. **FAC** - מתקנים (Facility) הם מבנים מעשה ידי אדם. לדוגמא: בתים, מפעלים, איצטדיונים, משרדים, מגדלים, בתי כלא, מוזיאונים, חניונים, נמלי תעופה, רחובות, כבישים, תחנות רכבת ואוטובוס, גשרים, מנהרות. בהכללה גסה, מתקנים הם אובייקטים שמטופלים בתחומי האדריכלות וההנדסה האזרחית.

שני קרונות ראשונים של ה [כרמלית] החדשה עושים דרכם...

f. **EVE** - אירועים (Event) בעלי שם שגור או מוגדר, אותו ניתן להכיר בצורה עצמאית גם מחוץ להקשר. לדוגמא, אירועי ספורט, פסטיבלים, מלחמות, קרבות.

מחר תערך ההגרלה שתקבע מי תארח את [מונדיאל 2026]

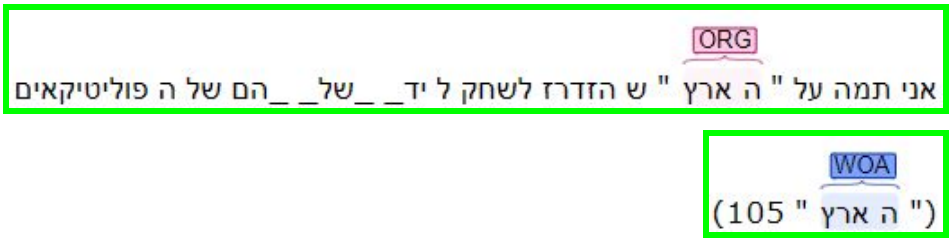
g. **DUC** - מוצרים (Product) בעלי שם: מכשירים אלקטרוניים, רכבים, נשקים, מוצרי מזון.

"[הדרימליינר]" בואינג [787] הוצג על ידי אל על... (היצרן וחברת התעופה יתוייגו כ-ORG)

h. **WOA** – יצירות אומנות (Work of art) בעלות שם שגור או מוגדר, אותו ניתן להכיר בצורה עצמאית גם מחוץ להקשר: ספרים, שירים, ציורים, פסלים, סרטים, סדרות טלויזיה, תיאוריות או עבודות מדעיות וכו'. לא מתייגים כותרות של מאמרים.

בסרט [סופרמן], קלארק קנט הוא סופרמן... ("סופרמן" ו-"קלארק קנט" יתוייגו PER)

WOA/ORG – כשהטקסט מתייחס לגיליון של עיתון, מתייגים WOA, כשההתייחסות היא למערכת העיתון או להנהלתה מתייגים ORG.



i. **ANG** – שפות (Language).

ל [ערבית] מעמד מיוחד במדינה...

מבחן השאלה[1]

הבדיקה הראשונה כדי לדעת האם ביטוי מועמד להיות שם-ישות, היא לבדוק האם הוא עונה על השאלה "מה *שם* הישות?" (What's the entity *name*?), או ליתר דיוק "מה *שם* ה-X?", כש-X הוא הקטגוריה המועמדת.

לדוגמא, במשפט "עשרות אנשים מגיעים מתאילנד...", הביטוי *תאילנד* מועמד להיות שם-ישות מהקטגוריה GPE. נשאל את השאלה "מה *שם* הישות הגיאו-פוליטית?", ונראה שאכן *תאילנד* היא תשובה טובה לשאלה. עבור PER "מה שם האיש/ה?", ORG – "מה שם הארגון?" וכו'.

מה לא מתייגים (כללי)

יש לשים לב שמבחן השאלה מדגיש את היות הביטוי שם. לעומת זאת, ביטויים העונים על השאלות "מה *סוג* הישות?", "מה *תפקיד* הישות?", ולא על מבחן השאלה – לא רלוונטיים לתיוג.

לא מתייגים כינויי גוף (היא, הן) וביטויים נומינאלים (שהם על פי רוב שמות עצם כלליים). ביטויים אלה אכן משמשים להצבעה על ישות יחודית בעזרת הקונטקסט, אך לא מצביעים עליה באופן עצמאי. "**הנשיא** נאם אתמול..." אין תיוג ל-"הנשיא". במבחן השאלה, הביטוי "הנשיא" לא יכול לענות על השאלה "מה שם האיש?".

יש לשים לב להבדל בין ביטוי נומינאלי לקיצורים שהם שמות. לדוגמא "ההסתדרות" הוא קיצור שהפך לשם שגור עבור "ההסתדרות הכללית של העובדים בארץ ישראל".

לא מתייגים ישויות שלא משתייכות לקטגוריות.

במשימה זו לא מתייגים שמות תואר (adjectives).

---

[1] ראו נספח בסוף מסמך זה לפירוט נוסף לגבי ההחלטה האם צירוף שמני (Noun Phrase) הוא שם, וכן קיצורי שמות.

דיוק התיוגים

לאחר שזוהה ביטוי המועמד להיות שם-ישות, התיוג נדרש לדייק בשני היבטים מרכזיים: שיוך הישות לקטגוריה המתאימה, וזיהוי גבולות שם הישות (היכן מתחיל והיכן מסתיים). נפרט לגבי כל אחד מההיבטים:

שיוך לקטגוריה המתאימה

כשישנה התלבטות בין שתי קטגוריות, כלל האצבע הראשון במשימה הוא *תיוג לפי משמעות*. מתייגות מתבקשות לשאול את עצמן תמיד: "בהינתן ההקשר, מי או מה *הישות המוצבעת* על ידי מחרוזת הטקסט הזו?" כלל זה עומד בעדיפות העליונה לאורך כל המשימה, ועל פיו תיבחר הקטגוריה. יש להשתמש בכל הקונטקסט הזמין, וניתן גם להיעזר בחיפוש באינטרנט, לדוגמא:

1. במשפט "במשחק הכדורגל בשבת [**סכנין**] ניצחה את [**בני יהודה**]", גם סכנין וגם בני יהודה יתוייגו בתור ORG, שכן בשתיהן ההצבעה היא על קבוצת ספורט (ועל אף ההצבעה הנפוצה יותר של המילה סכנין על GPE).

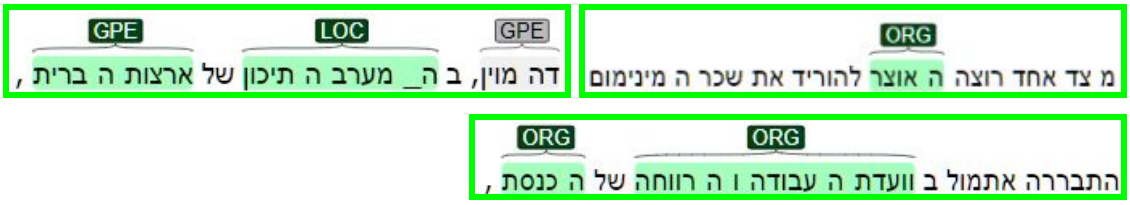
2. במשפט "הראשון לציון נאם אתמול בעצרת..." - אין תיוג לביטוי "הראשון לציון". הביטוי הוא תואר של כל רב ראשי ספרדי לישראל, ואינו כינוי או שם של רב מסויים. לא להתבייש לחפש כשלא בטוחים.

גבולות הביטוי

1. מיליות יחס תחיליות לא יתויגו כחלק מהשם - "מ **[תאילנד]** ל **[ישראל]**..."

2. ה' הידיעה תתויג כשהיא חלק מהשם - "**[ה ועדה ל אנרגיה אטומית]**", "**[ה כנסת]**"

   "ה [**אייפון**] החדש" (ה' הידיעה במקרה זה אינה חלק מהשם).
   יש לכלול את ה' הידיעה בתיוג כשהיא חלק מהשם. כשמדובר בביטוי מיודע המורכב מכמה מילים, תמיד מתייגים את ה' הידיעה הפותחת (לדוג' "המערב התיכון"). יש לשים לב לתקינות הביטוי.



3. ניתן לפרק ביטוי סמיכות ולתת תיוג שונה לכל חלק - "רחובות **[תל-אביב]**"

4. ביטויים שנחלקים על ידי מילות יחס (ב', מ', ל', על וכו'):

   a. שומרים על הנחיות העצמאות לביטוי, ואם יש צורך מתייגים באופן מקונן

      i. "**[הקרן החדשה לישראל]**" - תיוג מקונן

      ii. "**[משרד החוץ]** ב**[ירושלים]**" - תיוג נפרד

5. ביטויים שנחלקים על ידי מילת שייכות (של):

   a. על פי רוב יש לתייג בנפרד - "**[ועדת העבודה והרווחה]** של **[הכנסת]**"

   b. רק במקרים בהם בבירור לא ניתן להפריד מבלי לאבד את עצמאות הביטוי, התיוג לא יופרד - "**[הבית של [פיסטוק]]**"

6. קידומות של תארי כבוד, תפקידים ומאייכים אחרים, ככלל לא יתוייגו אלא אם יש בכך צורך הכרחי (כלל אצבע - *"תיוג מינימאלי, רק לא להגזים"*): הבדל בין:

a. תפקידים - "הנשיא **[מיטראן]**", "הסנאטור **[פיט וילסון]**", 'ח"כ **[אורה נמיר]**"

b. תארי כבוד - "מר **[כהן]**"

c. קידומות מיודעות - "השבועון **[לאקספרס]**", "העיר **[מיניאפוליס]**"

d. "שכונת **[בקעה]**", "מדינת [**נבראסקה]**"

e. מקרים בהם כן יכללו הקידומות:

i. הביטוי מאבד מהעצמאות שלו או משנה משמעות - "[**כביש באב אל-ואד]**", **"[שדרות רוטשילד]"**, **"[מלחמת ויאטנאם]"**

ii. בביטויים מיודעים (Definite Descriptions) שהם שמות הקידומות נוטות להיות יותר הכרחיות, ולכן על פי רוב כן יתויגו - **"[ארגון נפגעי המשכנתאות וחסרי הדיור]"**

7. שמות תואר - **"[מחלקת המדינה]** האמריקאית", **"[המערב התיכון]** האמריקאי"

8. קיצורים או שמות חלקיים -

a. ככלל, מתייגים, אלא אם כן מדובר בביטוי נומינאלי (ראה פירוט בהמשך) - למשל, כן מתייגים שמות פרטיים או שמות משפחה, קיצורים כמו "ההסתדרות", "האוצר" עבור "משרד האוצר", "הבקעה" עבור "בקעת הירדן", "השטחים" עבור "השטחים הפלסטיניים הכבושים" וכו'.

9. מרכאות יכללו בתיוג רק אם הן מופיעות באמצע השם - **"[ויטו "הסנדק" קורליאונה]"**

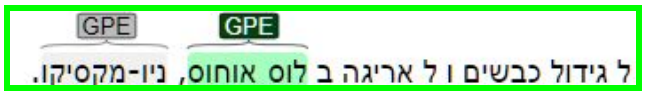
10. אליפסיס - השמט

a. במקרים של אליפסיס, כלומר, השמטה של מילה אחת או יותר, שאפשר להשלים על פי ההקשר, יש לפעול לפי הכללים הרגילים. במקרים בהם האיחוד לא משאיר ביטוי תקין, יש להפריד ולתייג בנפרד. לדוגמא:

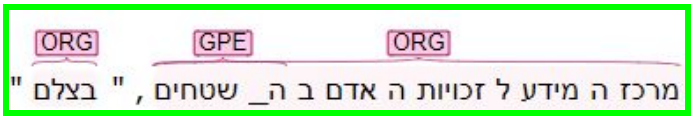
b. בדוגמא **"[ביל]** ו **[הילרי קלינטון]"** השם הפרטי ביל יתוייג כ-PER. השם השני נשאר תקין ובר תיוג כשהוא מאוחד לשם המשפחה ולכן הם יאוחדו ויתויגו יחדיו PER.

c. כמובן רק אם הביטוי היה ראוי לתיוג במקרה היה מופיע בנפרד לגמרי, לדוגמא "קרנות **[פורד]** ו **[רוקפלר]"** - לא מתייגים "קרנות פורד" כי הביטוי לא תקין.

11. רצפי ישויות מיקום - מתייגים בנפרד.



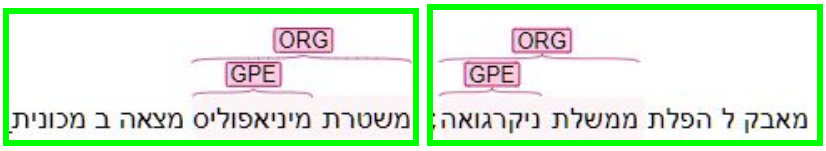
12. תמורה - כאשר מופיעים שני שמות שונים לאותה הישות ברצף, יש להפריד לשתי ישויות. כמובן שכל חלק יקבל תיוג רק הוא שם-ישות, ולא סתם תיאור.

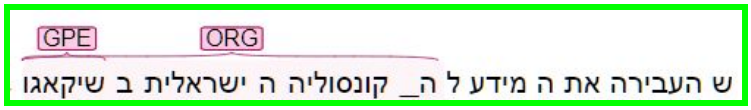


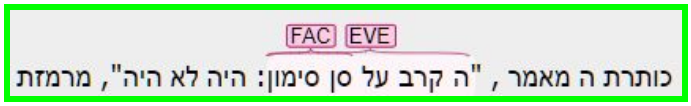
תיוג מקונן

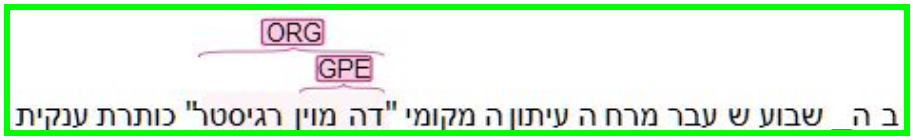
13. בישויות מקוננות, מתייגים תיוג מקונן. כל שם-ישות יקבל תיוג, גם אם הוא חלק משם ישות אחר. יש לשים לב לכל ההנחיות האחרות גם עבור הישויות החיצונית וגם הפנימיות.,

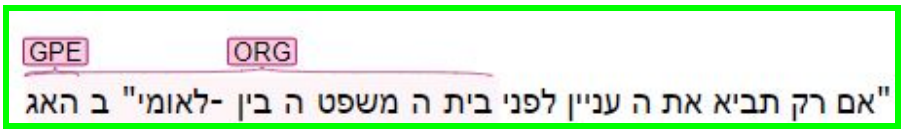
14. דוגמאות בהן ביטוי ראוי לתיוג מקונן:











a. ועידת המפלגה הדמוקרטית

b. האיסט סייד התחתית של ניו-יורק, תיאוריית היחסות של איינשטיין

c. אוניברסיטת קיימברידג/בוסטון/אריזונה

d. אירועי הר הבית

15. דוגמאות מתי ביטוי לא ראוי לתיוג מקונן:

a. קרוי על שם

   i. קרן פורד

   ii. פרס נובל

b. מחוז פריז, מדינת ניו-יורק

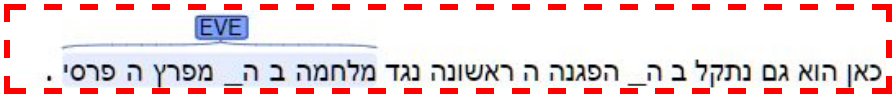
c. ירושלים המערבית? דרום לבנון?

נספח –

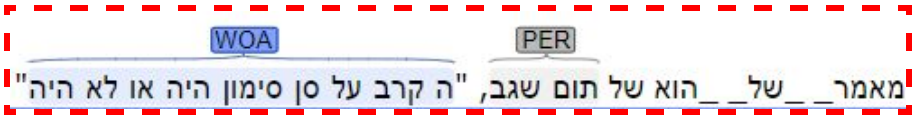
1. פירוט לגבי ההחלטה האם צירוף שמני (Noun Phrase) הוא שם

a. השאלה "האם המחרוזת היא ***שם***?" היא קריטית. מתייגים שמות, לא מתייגים כותרת או תיאורים, במיוחד ב-WOA, EVE. לא מתייגים כותרת של מאמר. מתייגים אזכורים של ***שמות*** אירועים בלבד ושל ***שמות*** יצירות בלבד. צריך להפעיל שיקול דעת, ואם לא בטוחים שזה שם, לא מתייגים.

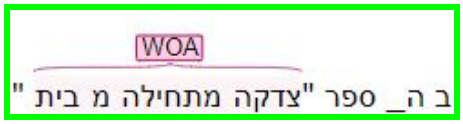
i. **תיוגים לא נכונים. לא שם של ישות:**

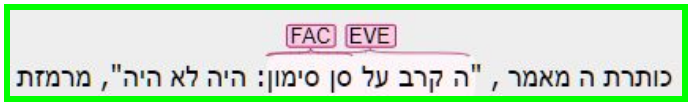


לא מתייגים *כותרת* של מאמר או כתבה.



ii. **תיוגים נכונים – שמות של יצירות ולא כותרות, שמות אירועים:**





b. **מתי ביטויים מיודעים הם גם שמות?**

i. **הבעיה**: כאשר מתייגים שמות עצם פרטיים (Proper Names/Nouns) הזיהוי והסיווג הוא יחסית פשוט.[2] אבל יש שמות עצם שהם לא כאלה, ורובם שייכים קטגוריה Definite Descriptions (תיאורים מיידעים): לדוגמא: "הילד של השכן" או "הבית הלבן".

ii. **הפתרון**: בביטויים אלה צריך לעשות את ההבחנה בין תיאורים לשמות. בכלל, יש ללכת גם כאן לפי מבחן השאלה (האם הביטוי עונה על השאלה "מה שם הישות?"), ולשים דגש על עצמאות ההצבעה לישות (שמות מצביעים באופן עצמאי על הישות, בעוד תיאורים מחייבים הקשר)

iii. **דוגמאות:**

1. ישנם מקרים ברורים, למשל תיאורים מהסוג "הילד של השכן", "הכלב שאכל את המחברת" לא יזכו לתיוג, לעומת שמות כמו *"הבית הלבן"* או *"הקרן*

---

[2] לא מצאתי תרגום תרגום ישיר בעברית למושג Proper Name (הבחנה בין Proper Name ו-Proper Noun). ההבחנה בין השניים היא על פי רוב כי Noun מתאר מילה אחת, בעוד Proper Name יכול להיות ביטוי המורכב מכמה מילים.

החדשה לישראל" שבבירור יתויגו. ישנן דוגמאות בהן זה פחות ברור. במקרים אלה יש להפעיל שיקול דעת, ניתן לחפש באינטרנט, ובכל מקרה כזה, אם לא בטוחים ברמה ודאות גבוהה, לא מתייגים:

2. *"הקונסוליה הישראלית בשיקאגו"* - מתייגים. עונה היטב על השאלה "מה שם הארגון?". זהו שם הישות. אין לה שם אחר. כך גם ברישום שלה באתר משרד החוץ ובאתר הרשמי שלה.
3. *"בית הספר בנווה-מונוסון"* - לא מתייגים, למרות שמתייחס לישות ספציפית יחידה, זה אינו שם בית הספר (בניגוד למקרים כמו *"בית הספר עירוני ה' בתל אביב"*).

c. צריך לעשות את אותה ההבחנה בין (Proper Names ו-Definite Descriptions) גם בתיוג של **קיצורי שמות:**

i. Proper Names שהם שמות חלקיים או מקוצרים, כמו למשל השימוש בשם הפרטי של אדם, ראשי תיבות של מיקום, כינויי חיבה ("ביבי", "בוז'י") תמיד יתויגו כישות מהסוג מתאים.

ii. קיצורים של Definite Descriptions אינם שקולים לשמות מקוצרים, ולא יתויגו כישות.

1. **הסבר**: ביטויים מיודעים (definite descriptions) לרוב מכילים שם עצם כללי מיודע ולאחריו אפיון יותר ספציפי , לדוגמא: "שירות התעסוקה" "התאחדות האיכרים". בהמשך הטקסט ייתכן שימוש בשם העצם הכללי בלבד (לדוגמא: "השירות" "ההתאחדות"), לאזכור הישות אחרי שהיא הוזכרה במלואה . במקרים אלה **לא יתבצע תיוג**, כפי שלא מתבצע תיוג של כינויי גוף (היא, הוא...):
   a. "ועדת העבודה והרווחה של הכנסת" ← "הועדה"
   b. "מנזר סן סימון" ← "המנזר"
   c. "שירות התעסוקה" ← "השירות"
   d. "התאחדות האיכרים" ← "ההתאחדות"
2. ***במקרים נדירים*** יותר הקיצור הזה הוא גם שם של ממש. קיצור זה יהיה בשימוש נפוץ בצורה עצמאית, גם ללא אזכור מוקדם או קונטקסט ברור. במקרים כאלה לתייג רק כשיש ביטחון מוחלט בכך שזהו שם שגור ולא אזכור הקשר מקומי/משפטים קודמים. לדוגמא:
   a. "המוסד למודיעין ולתפקידים מיוחדים" ← "המוסד"
   b. "ההסתדרות הכללית של העובדים בארץ ישראל" ← "ההסתדרות"