הנחיות תיוג איגוד אזכורים בעברית

Hebrew Coreference Resolution Guidelines

רפאל שקד גרינפלד, רעות צרפתי
shakedgreenfeld@gmail.com | reut.tsarfaty@biu.ac.il

המשימה

המשימה היא איגוד אזכורים (Coreference Resolution), קישור כל האזכורים הספציפיים בטקסט המצביעים לאותם ישויות ואירועים. אזכור יהיה מוגבל לצירופים שמניים (Noun Phrases), מבני סמיכות (סגל הבית הלבן) ותואר הפועל (adverbs) בהתייחס לזמן (אתמול) ומקום (כאן), כאשר כל אזכור אשר באותו האגד מתייחס לאותה ישות.

סגמנטציה

הסגמנטציה של הטקסט (חלוקת הטקסט למילים) תהיה חלוקה ליחידות הבסיס הקטנות ביותר הנושאות משמעות. כך שהחלוקה למילים עבור המשפט.

ח"כ אלי דיין (מערך) הגיש הצעת חוק שלפיה יוטל היטל על מעסיקי עובדים זרים, כדי למנוע העדפתם על עובדים ישראליים.

תהיה:

*ח"כ, אלי, דיין, (, מערך,), הגיש, הצעת, חוק, ש, **לפי_**, **_היא**, יוטל, היטל, על, מעסיקי, עובדים, זרים, , ,כדי, למנוע, **העדפה**, **_של_**, **_הם**, על, עובדים, ישראליים .*

נשים לב כי כל מילה מחולקת עד לרמת המורפמה, היחידה הלשונית הקטנה ביותר הנושאת משמעות, כך שהמילה "לפיה" תתפרק לשתי המורפמות "לפי_" ו"_היא" כאשר הסימן "_" מחווה על כך שהמילה היא למעשה מורפמה שהתפרקה.

אזכורים

אזכורים הם היחידות אשר מרכיבות את אגד האזכורים. אזכור הוא ביטוי המתייחס לאיזה שהיא ישות (named entity and non-named entity) שניתן להצביע עליה במציאות היום יום, בין אם היא מתייחסת לאיש ספציפי, קבוצת אנשים, עצם דומם או ישיויות מופשטות.

דוגמאות לאזכורים (לא כל האזכורים בטקסט הבא מסומנים):

[ח"כ אלי דיין] הגיש [הצעת חוק] שלפיה יוטל [היטל על [מעסיקי עובדים זרים]], כדי למנוע העדפתם על עובדים ישראליים. הוא צפוי לקדם [מס זה] במושב הקרוב אחרי שיחזור מ[החופש שלקח]

ח"כ אלי דיין, הוא - איש ספציפי
מעסיקי עובדים זרים - קבוצת אנשים
הצעת חוק, מס זה, החופש שלקח - ישיויות מופשטות

במשימת התיוג, האזכורים יחולצו באופן אוטומתי אך היות והחילוץ אינו מושלם, ומספר אזכורים יהיו חסרים/פגומים יש צורך לסמן את החסרים בהתאם להנחיות ולהגדרה של "מה הוא אזכור". מצד שני, חלק מאותם מקטעים שחולצו אינם נופלים תחת ההגדרה ואין לסמנם כלל גם לא בקבוצת אזכורים יחידנית (קבוצת אזכורים בגודל אחד) ויסומנו כאידיאומטי (idiomatic). כל האזכורים התקינים ישמשו בשלב השני כמועמדים לאיגוד האזכורים.

בחלק זה נגדיר איזה מהביטויים נופל תחת הגדרת אזכור ואיזה לא.

צירוף שמני

צירוף שמני (Noun Phrase) מורכב משם עצם (Noun) או שם גוף (pronoun) והמילים המחוברות אליהן בעץ המבנה התחבירי ויוצגו למתייגים כמועמדים לאגד אזכורים (coreference), כאשר ככלל, הצירופים יחולצו באופן אוטומטי מעץ התחביר (אם המסמך מתויג כחלק מהHTB) או באמצעות מודל שאומן לביצוע משימה זאת.

מילות יחס

מילות יחס לא יהוו חלק מן הצירוף השמני, לדוגמא במשפט: *[ה מלך] חי ב [ארמון]* אנו לא נכליל את מילת היחס ב (in) כחלק מהאזכור ארמון, היות ואינה חלק אינהרנטי מהצירוף. לחילופין במשפט *ב[מלון ברמת עמון]x, אפשר לקנות קילו ענבים וחצי מלון, נורא זול שם.*
האזכור "מלון ברמת עמון" הוא יחידה לוגית אחת ולפיכך כן נכליל את ה "ב" השנייה כי היא כן חלק מכלל הצירוף, כאשר היות וה "ב" הראשונה, אינה מהווה חלק מהאזכור נשמיט לא נסמן אותה כחלק מהצירוף השמני.

תמורות אל מול יחס אנאפורי

התמורה היא סוג של לוואי, אשר צירופו לגרעין אינו יוצר סמיכות; חלק ממשפט, המהווה ביטוי נרדף, או ביטוי נוסף לתיאור אותו הדבר. כאשר נסמן כאזכור, אנו מחד לא נפריד את תיוג הצירוף השמני בין צירוף שהוא חלק מיחס אנאפורי לבין צירוף שהוא תמורה ביחס לגרעין ונתייגם באופן זהה ומאידך, אך אנו נסמן את כלל מרכיבי המשפט: הגרעין, התמורות ואת כלל הצירוף השמני.

מאמצים דיפלומטיים לשחרור [ישראלי שכלוא בטייוואן]: "[הוא] חלש וחולה"

[[הדוקטור], [אבי כהן]], הצליח לרפא אותי.

[[ראש הממשלה], [בנימין נתניהו]], נאם אתמול בכיכר.

צירופים מקוננים

בשונה מ Base NP Chunking, המוגדר ע"י חלוקה של טקסט למקטעים לא חופפים וסימונם של הצירופים השמניים הלא רקורסיביים

[the nation] 's [manufacturing titans] typically jet off to [the sunny confines] of [resort towns] like [Boca Raton] and [Hot Springs].

אולם עולה [כאן] [השאלה], כיצד הסכימו [המנגנונים] ו[מושלי המחוזות] מלכתחילה הצגת מועמדות של [הם] של [ראשי כנופיות] ו[סוחרי סמים], כדי שייצגו את [מפלגת הרוב].

וגם בשונה , מצירופים שמניים כלליים

[the nation's manufacturing titans] typically jet off to [the sunny confines of resort towns like Boca Raton and Hot Springs].

אולם עולה [כאן] [השאלה], כיצד הסכימו [המנגנונים ומושלי המחוזות] מלכתחילה [הצגת מועמדות של הם של ראשי כנופיות וסוחרי סמים], כדי שייצגו את [מפלגת הרוב].

אנו מאפשרים סימון של כלל הצירופים היכולים להתייחס ליישות ספציפית בעולם ובפרט מאפשרים צירופים מקוננים.

[[the nation] 's [manufacturing titans]] typically jet off to [[the sunny confines] of [resort towns] like [Boca Raton] and [Hot Springs]].

אולם עולה כאן השאלה, כיצד הסכימו [[המנגנונים] ו[מושלי המחוזות]] מלכתחילה [הצגת מועמדות של [הם] של [[ראשי כנופיות] ו[סוחרי סמים]]], כדי שייצגו את מפלגת הרוב.

יוצאי דופן לכלל דנן, הם מבני סמיכות וכינויי הדגמה (Demonstrative pronoun)

מבני סמיכות מכוננים

למרות שבעברית לא ניתן לפרק מבנה סמיכות ולשמור על המשמעות הסמנטית של הביטוי, אנו מחד לא נפרק מבנה סמיכות לכמה אזכורים נפרדים אך מאידך נאפשר מבנה מכונן. כמו זאת יש לציין, כי מבנה סמיכות יכול להוות חלק מצירוף שמני מכונן אחר.

[עורך [ה עיתון "אל - אחאבר"]] תוקף על רקע זה את [ראשי [המפלגה]], אף שביטלה את מועמודת

כינויי הדגמה מכוננים

כינוי הדגמה הם כינויים שבאים לסמן לנמען דבר מסויים. מילים כגון *זה* או *אלה* נחשבות מילות הדגמה. דוגמאות לביטויים עם כינויי הדגמה יהיו: *רקע זה; מנהיגים אלה; וועדה זו; פעילות פוליטית כלשהיא*

בשונה ממבני סמיכות, שם אנו מאפשרים מבנה מכונן,עבור כינויי הדגמה (dem) אנו לא נאפשר כינון של הביטוי. למרות שחלקי הדיבר של הביטוים הנ"ל הם בצורת NOUN PRON, כל מילה אינה יכולה לעמוד בפני עצמה ועל כן אינן ניתנות לפירוק ללא איבוד משמעות.

[הן] גם אינן אמורות לקחת [חלק] ב[פעילות פוליטית כלשהי].

ביטויים מכמתים

ביטויים מכמתים לסוגיהן של ישות X שמשתנה ע"י מכמת Y כמו ביטויים קרדינליים (עשר, מיליארד ,אלפים), חלקיים (כמה, חצי, מעטים) או מידות (ליטר, קוב) ותיאורים מקבצים (עדר, צבא)

שלושה אנשים - קרדינליים
הרבה חיות - חלקיים
מקטע של כביש 6 - מידות
נחיל דגי דקר - מקצבים

יסומנו גם כיחידה אחת אך גם כביטוי מכונן, משמע, נסמן את כלל הצירוף השמני מחד ומאידך נסמן את הצירוף הפנימי.

[*שלושה [אנשים]] - קרדינליים*
[הרבה [חיות]] - חלקים
[מקטע של [כביש 6]] - מידות
[נחיל [דגי דקר]] – מקצבים

שמות תואר

שמות תואר לא יסומנו כאזכורים נפרדים
הילד היפה
יסומן כאזכור יחיד [הילד היפה] כאשר שם התואר "היפה" יהיה חלק מהאזכור "הילד"

תואר הפועל

תוארי הפועל ככלל לא יסומנו כאזכורים (למעט מקרים של זמן ומקום שיפורטו בהמשך).

במידה ויופיע תואר הפועל לאחר מציאת שורש אזכור, התואר יהווה חלק מהאזכור בסיס. לדוגמא הביטוי "יותר" כמו "אנשים יפים יותר" ,או "מאוד" כמו "הילד היפה מאוד" יהיו חלק מהאזכור.

למעט ביטויים של תואר הפועל, המתייחסים לזמן ולמקום יסומנו גם כן כאזכור, היות ויכולים להוות חלק מאגד אזכורים. עבור זמן, כל ביטוי המתייחס לזמן מוגדר יסומנו כאזכור: *אתמול, שלשום, אמש ,מחר, אז פעם, [ה]שנה ,[ה]חודש, [ה]שבוע, [ה]יום.*

באופן דומה נסמן את תארי הפועל של מקום:
פה, כאן, שם, למעלה, למטה, קדימה, אחורה, בצד, מזרח, מערב, צפון, דרום.
כל ביטוי עם יחס סמיכות

תואר פועל של זמן או מקום שאינו חלק מרשימה סגורה זו, יסומן באופן ידני.

חברות המעסיקות עובדים זרים זוכות במכרזים, היות והן מציעות [שירותים זולים יותר]

*הוא רואה במס מעסיקים **מס מעוות**.*

"מס מעוות" הוא שם תואר למס מעסיקים, לכן **לא יסומן כאזכור**

אמות מידה

אמות מידה שמופיעות בפני עצמן, לא יסומנו כאזכורים. במידה ומתארים דבר מה באמצעות אמת מידה של גודל, זמן או מרחק אנו לא נסמן את אמת המידה כאזכור. לדוגמא אם ידובר על כך שהעונש על רצח הוא *כעשרים שנים*, או לחילופין מישהו מתאר שהוא הולך כל יום *30 ק"מ או שמישהו יותר גבוה מעץ אלון*. אנו לא נסמן את "עשרים שנים", "30 ק"מ" ו "עץ אלון"

רנדי בארנס עלול להפסיד יותר מחצי מיליון דולר.

שיאן העולם באכילה מהירה שוקל יותר מ400 קילוגרם

המגלשה הזאת ענקית! היא יותר גבוהה ממנוף!

{חצי מיליון דולר, 400 קילוגרם, מנוף} לא יסומנו כאזכורים וייחשבו כאידיאומטים.

במידה ואמת המידה היא חלק מהאזכור, או משמשת כמילה נרדפת לשם עצם, אנו **כן** נסמנה
השיאן העולמי ב[400 מטר] הוא מייקל ג'ונסון

400 מטר כן יסומן כאזכור - היות ו400 מטר בעצם משמש כביטוי נרדף ל "תחרות ריצת 400 מטר"

מבני סמיכות

בעברית קיימת תופעה תחבירית ייחודית של מבנה סמיכות, המחברת בין שני שמות עצם הקיים ביניהם קשר הדוק ("הבית של הנשיא", "בית הנשיא", "ראש ההר","עורך העיתון" וכו') במבנה סמיכות שכזה אנו נסמן את כלל מבנה הסמיכות כצירוף שמני אחד.
[עורך ה עיתון "אל - אחאבר"] תוקף על רקע זה את [ראשי המפלגה], אף שביטלה את מועמדתם

אנו לא נפרק מבנה סמיכות בכדי לא ליצור ביטויים חסרי משמעות סמנטית. מצד שני, המידה ומבנה הסמיכות מורכב מיותר משני מילים, כן נסמן את הביטוי המקונן כאזכור.

possessive, ביטויי שייכות

במקרה של ביטויי שייכות באופן דומה למבנה סמיכות, לא נרצה לפרקם כי נאבד את משמעות הביטוי המקורי ונסמן גם את הביטוי המכונן.
לדוגמא את הביטוי:
סיורים של צה"ל

נסמן בצורה הבאה:
בגזרת עזה נערכים מדי יום [סיורים של [צה"ל]], הם נערכים בשעות הערב והבוקר

במסגרת חוק זה אנו נתמוך בביטויים כמו:

1. סיורים של צה"ל
2. הכוונה המקורית של משה
3. חלקים גדולים למדי של החברה
4. רגעי החסד הגדולים שלו (של_ _הוא)
5. התוכנית הראשונה שלהם (של_ _הם)
6. הקומה השנייה של הבניין

פעלים VERBS

נאפשר פסוקיות NP המכילות פסוקיות VP המתארות את שם העצם
לדוגמא: *חברות המעסיקות עובדים זרים*
נסמן בצורה הזאת: *[[חברות המעסיקות [עובדים זרים]]*
ולא בצורה הזאת: *[חברות] המעסיקות [עובדים זרים]*

מצד שני, אין אנו נסמן פועל עם הטיה בינונית.
[אנשים אלה] למעשה משמשים [עובדים שכירים זולים].
הפועל *"משמשים"* לא יסומן כאזכור.

שמות פעולה

שם הפעולה הוא שם שנוצר מפועל, אך אין לו זמן וגוף. שמות הפעולה מציינים לרוב פעולה, תהליך מצב או תוצאה. (הלכתי - שורש הלך *גוף ראשון זמן עבר ושם הפעולה* הליכה). אנו **נסמן כאזכור** את כלל שמות הפעולה[1] .

אם אמנם יאלץ להיעדר מפעילות שנתיים בגלל השעיה על [שימוש בסמים]

אנו **נסמן** *"שימוש בסמים"* כאזכור

[1] מידע נוסף על שמות פעולה:
https://meyda.education.gov.il/files/Pop/0files/ivrit_habaah_lashon/chativat_elyona/on-line-learning/shem-poal-peula.pdf

אני אוהב לרקוד [ריקודי עם] עם אשתי

אנו **נסמן** *"ריקודי עם"* כאזכור

סימני פיסוק

ככלל אם כחלק מהאזכור מופיע סימן פיסוק המצריך סימטריה (לדוגמא: גרש, סוגריים), אנו נסמן את האזכור בצורה השומרת על הסימטריה.

"עורך הדין" ***ולא*** *עורך הדין"*
"גיפריס" ***ולא*** *"גיפריס*
(הארץ) ***ולא*** *(הארץ*

במקרה והאזכור המוצע אינו תקין יש לסמנו כidiomatic ובנוסף לסמן אזכור חדש אשר כולל את גם את הגרש הפותח וגם את הסוגר או לחילופין את זוג הסוגרים.

ביטויים אידיאומטיים אשר אינם נחשבים לאזכור Idiomatic Phrases

ביטויים

במקרה ושם העצם משמש כביטוי או חלק מביטוי אנו לא נסמנו כאזכור. לדוגמא, הביטויים "עם זאת", "על כך", "על פי" לא יסומנו כאזכורים וגם המילים "פי" "זאת" ו"כך" לא יסומנו כאזכורים.

הילדים שוב הפריעו בכיתה ועולם כמנהגו נוהג, על פי המורה, החומר בכל אופן הועבר בהצלחה.

שמות עצם אשר אין להם התייחסות בטקסט

ביטוי אידיאומי הוא ביטוי המועמד להיות אזכור אך אינו אזכור. מועמד ייחשב אידיאומטי אם לא יתייחסו לביטוי המסומן בטקסט או לחילופין.

לדוגמא:

שחקני [בית"ר [ירושלים]] חזרו ארצה למשחק הבית ב[עיר הבירה].

האזכור ירושלים ייחשב כביטוי אידאומטי היות והכותב התייחס לקבוצה בית"ר ירושלים ולא לעיר אשר שמה הוא חלק משם הקבוצה. לכן במקרה דנן, לא רק שהביטויים "עיר הבירה" ו "ירושלים" לא יהיו תחת אותו *האגד*, הביטוי ירושלים אף ייחשב לאידיאומטי.

מצד שני,

ביטויים כמו "ממשלת ישראל" מבטאים מבנה הדומה למבנה שייכות (possesive) ויכולים להקרא כ "הממשלה של ישראל" (לא כמו בית"ר ירושלים שאינה שקולה ל *הבית"ר של ירושלים*) ולכן ישראל תחשב כאזכור תקין **ולא** כאידאומטי.

ביטויים פחות ספציפיים

במידה ובין המועמדים לאזכורים ישנם מספר מקטעים אשר קיים ביניהם יחס הכלה והם מתייחסים לאותו אזכור, אנו נסמן כאזכור את הביטוי הכי ספציפי (לרוב הוא יהיה הביטוי הארוך ביותר) אשר מתייחס לדבר מה. כל שאר הביטויים יסומנו כאדיאומטיים.

[[נשיא מועדון השחמט] שנבחר כבר 5 פעמים ברציפות] מתמודד על ראשות המועדון זה הפעם השישית

היות ו"נשיא מועדון השחמט" הוא ביטוי פחות ספציפי מהביטוי "נשיא מועדון השחמט שנבחר כבר 5 פעמים ברציפות" אנו נסמן:

נשיא מועדון השחמט - אידיאומטי
נשיא מועדון השחמט שנבחר כבר 5 פעמים ברציפות - אזכור תקין

ברור ש_אנחנו עדים באחרונה ל[[התפתחויות שלא הכרנו בה_עבר], בה_עולם המוסלמי הקיצוני] שמשפיעות עלינו מאד]

בדוגמא הנ"ל קיים האזכור "התפתחויות שלא הכרנו בה_עבר, בה_עולם המוסלמי הקיצוני שמשפיעות עלינו מאד" שהוא הביטוי הספציפי ביותר המתאר את ה"התפחויות" ולפיכך נסמנו כאזכור תקין ואת שאר האזכורים כאידיאומטיים:

התפתחויות שלא הכרנו בה_עבר - אידאומטי
התפתחויות שלא הכרנו בה_עבר, בה_עולם המוסלמי הקיצוני - אידאומטי
התפתחויות שלא הכרנו בה_עבר, בה_עולם המוסלמי הקיצוני שמשפיעות עלינו מאד - אזכור תקין

כמו כן, גם במקרים בהם התיאור מופרד ע"י פסיק (,) אנו נחברו לראש הביאוי

[יונתן, בן קיבוץ נחשולים], הלך אתמול לגן

יונתן - אידיאומטי
בן קיבוץ נחשולים - אידיאומטי
יונתן, בן קיבוץ נחשולים - אזכור תקין

שמות עצם פרטיים

שמות לא יפורקו ונתייחס אליהם כמקשה אחת, בתוך זה נכלול תארים כגון "מר" "ד"ר" וכדומה אשר לא יופרדו משם העצם שהוא שורשם. אי הפירוק כולל בתוכו גם שמות של מקומות, שירים, ארגונים וכו'.

[פרופסור [דני אורבך]] התארח הערב בתוכנית פגוש את העיתונות.

בשיר [בית המלון ב[קליפורניה]] שרה להקת הנשרים על מלון בקליפורניה.

אני כבר שנים לקוח ב[בנק [ירושלים]] מאז שגרתי בבירת ישראל.

האזכורים שנסמן כתקינים הם:
- פרופסור דני אורבך
- בית המלון בקליפורניה
- בנק ירושלים

ונסמן כאידיאומטיים את הביטויים:
- דני אורבך
- קלפורניה
- ירושלים

טעויות בטקסט (ג״ביריש)

אם מופיע רצף תווים אשר אינו מילה בעברית ולא ניתן להבינה ולאחריה רצף מילים תקין אשר נראה כמו מילה תקינה, אנו נסמן רק את המילה התקינה כאזכור

קפידבינהמלך קפינדרה הסכים לבטל את החוקה

אנו נסמן רק את *״המלך קפינדרה״* כאזכור (ולא את *קפידבינהמלך קפינדרה)*

תוצאות ופרטים עוקבים

אם מופיעות תוצאות של משחקים או איזה שהוא פרט לגבי שורש האזכור שמוסיף עליו פרטים, נסמן גם את הפרטים במידה והם קיצור של אזכור בפני עצמו

קייזרסלאוטרן (61 נק) מחכה לדיסלדורף (מקום 21)

אנו נסמן

[קייזרסלאוטרן ([61 נק])] מחכה ל[דיסלדורף ([מקום 21])]

סימון אזכורים

למרות שרוב האזכורים הפוטנציאלים מסומנים אפריורית באופן אוטומטי, קיימים מקרים בהם חלק מהאזכורים לא מוצעים בתור מועמדים בטקסט. במקרה כזה, יש לסמן את האזכורים ידנית.
בעת הסימון יש לשים לב לדגשים הבאים:

- לאחר סימון האזכור לא ניתן למחוקו. במידה והחלטתם לאחר מעשה שהאזכור שגוי, ניתן לסמנו כ idiomatic
- אין לסמן אזכור שכבר סומן.
- אזכור שצריך להיות מסומן הוא כל רצף מילים אשר נופלים תחת ההגדרה של אזכור תקין ואשר לא חולצו באופן אוטומתי כאזכור.
- אזכור לעולם יהיה באותו המשפט בלבד. לא נסמן רצף מילים אשר חוצה משפטים.

איגוד אזכורים

לאחר שהוגדרו האזכורים הפונטציאלים, בחלק זה אנו נגדיר את האופן בו נאגדם לכדי אגד אזכורים (או שרשרת אזכורים). בחלק זה נגדיר אילו אזכורים יאוגדו יחדיו ונפרט את סוגי האגדים השונים. כמו כן נדגים באמצעות דוגמאות מייצגות את המקרים השונים על מנת שישרתו את התיוג בצורה המיטבית.

קשר זהות

כינויי שם, כינויי גוף, שמות עצם מורכבים, תארי פועל וכינויי הדגמה אשר מתייחסים לאותה יישות, מאורע או מושג יקושרו יחדיו לאותו אגד אזכורים בקשר זהות. כל עוד אותם אזכורים מתייחסים לאותה יישות מאורע או מושג, אין הגבלה על הקשר.

1. *היתרון העצום של [ויליאמס_1] בסקרים פחת והלך, ו[הרפובליקאים] מרטו בייאוש את שערותיהם. [הם] התחננו לפני [ויליאמס_2] לשתוק.*
   a. *ה יתרון ה עצום של [ויליאמס_1] ב ה_ סקרים פחת ו הלך, ו [ה רפובליקאים] מרטו ב ייאוש את שערותיהם. [הם] התחננו לפני [ויליאמס_2] לשתוק.*

אגד אזכורים 1: {ויליאמס_1, ויליאמס_2}
אגד אזכורים 2: {הרפובליקאים, הם}

תמורות

קשר תמורתי הוא רצף אזכורים של ביטויים שמניים המופרד לרוב על ידי פסיק, מקף או סוגריים המתאר את אותה יישות . בפרוייקט זה אנו, ראשית נסמן את הביטוי הרציף הארוך ביותר כאזכור אחד ואין אנו נעשה הפרדה בין סוגי קשר שונים, כמו כן, נסמן כל חלק בתמורה בנפרד ונוסיפה לאגד.

איש ב טקסס לא פיקפק ש[[קלייטון ויליאמס], [חוואי ואיש נפט]], יביס אותה בקלות. [הוא] הופיע בתשדירי הבחירות שלו רכוב על סוס, עם מגבעת רחבת תיתורת, ופרט על נימי המאציזמו הטקסני.

אגד אזכורים: { "קלייטון ויליאמס, חוואי ואיש נפט" , "קלייטון ויליאמס", "חוואי ואיש נפט" , "הוא"}

שמות עצם מורכבים

במידה וחלה טעות בחלקה הראשון של המשימה, ועבור ביטוי מסויים יופיעו לנו תתי אזכורים נוספים שהם פחות ספציפיים אנו נאגד את הביטוי בעל המופע הספציפי ביותר לביטויים האחרים המפנים לאותה יישות. ואת הביטוי הפחות ספציפי אנו נשים באגד נפרד יחידי ו**לא עם האזכורים הנכונים**.

תופעה זו התבררה אתמול ב[[וועדת העבודה והרווחה של הכנסת], שדנה בנושא העסקת עובדים זרים]1. יו"ר [הוועדה]x, ח"כ אורה נמיר (מערך), טענה כי "מביאים עובדים זרים לישראל על תקן של מתנדבים מתאילנד, רק כדי לא לשלם להם שכר מינימום.

אגד אזכורים: { "וועדת העבודה והרווחה של הכנסת, שדנה בנושא העסקת עובדים זרים", "הוועדה"}
אגד יחידני: {" וועדת העבודה והרווחה של הכנסת"}

היות ו "*וועדת העבודה והרווחה של הכנסת*" הוא מופע פחות ספציפי מ "*וועדת העבודה והרווחה של הכנסת, שדנה בנושא העסקת עובדים זרים*" אנו נבחר לאגד את השני כחלק מהאגד ולא את הראשון. כמו כן, החלק הראשון, הפחות ספציפי ייחשב כ**ביטוי אידיאומטי ונסמנו באגד נפרד - שאף ביטוי אחר אינו מצביע אליו.**

כינויי גוף וכינויי רמז

בתוך ציטוט: כינויי גוף ורמז יקושרו אל הנושא אליו הם מתייחסים, גם במקרים בהם הכינוי הוא חלק מציטוט.

כפי שאמר [חוסה מרטי] לאחר שהותו בארה"ב "חייתי בתוך המפלצת ו[אני] יודע זאת אפילו מעבר לקרביו"

אגד אזכורים: { *חוסה מרטי*, אני}

ביטויים המופיעים כתוצאה של טוקניזציה:
כאשר הכינוי הוא תוצר של טוקניזציה,משמע הכינוי (pronoun) לא הופיע בטקסט המקורי והתווסף כתוצאה מהטוקניזציה אנו נסמן רק את הכינוי עצמו ולא את כל הביטוי כאזכור אשר מקושר אל הראש שלו.

משפט מקורי:

הם עודדו חברות לתעל את הפילנטרופיה שלהן
משפט לאחר טוקניזציה:
הם עודדו [חברות] לתעל את הפילנטרופיה של [הן]
אגד אזכורים: {חברות, הן}

בדוגמא זאת אנו לא נסמן את "שלהן" אלא רק את "הן" כביטוי אשר מקושר ל"חברות"

ביטויים אקספלטיביים (non-referential pronouns): כמו כן, כינויים אקספלטיביים שאינם מביעים משמעות או לא תורמים לה ישירות, לא יסומנו גם כן. אנו נזהה כינויים אלה באמצעות מבחן ההחלפה. אם לא ניתן להחליף את הכינוי בביטוי שמני (NP chunk) אז זהו ביטוי אקספלטיבי

[זה]_1 נראה שחם היום. [זה]_2 מקשה על הנשימה.

את "זה_1" לא ניתן להחליף בביטוי אחר, לעומתו את "זה_2" ניתן להחליף בביטוי "המסיכה שלי" (או המילה "חום"). לפיכך:

זה_1 - idiomatic
זה_2 - יכול להיות מקושר

אוגד לעומת כינוי גוף
במשפטים שמניים, כלומר משפטים שאין בהם פועל, יש בין שני חלקי המשפט העיקריים כינוי גוף נסתר: "הוּא" (ולפי הצורך הִיא, הֵם, הֵן). כינוי זה נקרא 'אוגד' על שם מעמדו כמעין איבר מחבר בין הנושא לנשוא. כמו כן, כינוי הרמז זֶה משמש לעתים גם כן בתפקיד אוגד.[2]

*הוראה **זה** מקצוע מאתגר*
*כדורגל **הוא** הספורט הכי פופולארי בעולם*

מילות אוגד **לא ייחשבו כאזכורים ככלל** ולא יהיו חלק משרשרת אזכורים בפרט
[דני] הוא_1 [האח של [יוסי]]. [הוא_2] היום בן 11.

אגד אזכורים: { דני, האח של יוסי, הוא_2}

מבנים קופולריים

קופולה הוא פועל "ריק" למשפטים בהם הפרדיקט הוא לא פעלי. הם יורכבו מהנושא, תכונה של הנושא וקופולה (במקרים רבים הקופולה תהיה אוגד) שתקשר בין הנושא לבין התכונה שלו. במקרים אלה, אנו לא נסמן את התכונה כאזכור של הנושא.

[דנית] הייתה [רקדנית בלהקת המחול של מוסקבה].

- *"דנית"* (הנושא) מקושרת באמצעות הפועל *"הייתה"* (הקופולה) ל*"רקדנית בלהקת המחול של מוסקבה"* (התכונה) ולכן התכונה לא תהיה מרופררת לדנית.

[2] האקדמיה ללשון עברית, /זה-בתפקיד-אוגד/https://hebrew-academy.org.il/2020/09/06

- היות ותכונה משמשת פה כתואר - *"רקדנית בלהקת המחול של מוסקבה" לא תסומן כאזכור כלל.*

מצד שני אם הקופולה מציינת תיאור חד חד ערכי לנושא, אנו כן נאגדם יחדיו.
[בנימין נתניהו] הוא [ראש ממשלת ישראל].

- *היות והביטוי "ראש ממשלת ישראל" מתייחס לבנימין נתניהו באופן בלעדי בקונטקסט הנ"ל, ואין הוא מתייחס לתכונה של בנימין נתניהו אנו כן נאגדם יחדיו*

ביטויים כלליים (Generic Mentions)

ביטויים כלליים, אשר אינם מכוונים לאף יישות ספציפית, בין אם הם אבסטרקטיים ובין אם הם ביטויים מכלילים יהיו מקושרים רק בקשר אנאפורי (או קטאפורי) לכינוי הגוף או לביטוי מיודע, אך לא יקושרו לביטויים כלליים אחרים ללא קשר שכזה.

ביטויים בצורת רבים כמו "בכירים", "ילדים", "אריות" וכדומה, יסווגו כביטויים כלליים ולפיכך יקושרו רק לכינויי גוף בעלי קשר אנאפורי. מכיוון שאין אנו נקשר ביטוי כללי אחד למשנהו, יהיה עלינו לפתוח אגד אזכורים חדש.

[ילדים] אומרים ש[הם] אוהבים את החופש הגדול.
אגד אזכורים: {ילדים (ביטוי כללי), הם}

מכל החיות אני הכי אוהב נחשים. חיות אלה מרגשות אותי מאד. הם מאד יפות.
אגד אזכורים: {נחשים (ביטוי כללי), הם (כינוי גוף עם אנפורי), חיות אלה (ביטוי מיודע) }

[הורים]1 צריכים להיות מעורבים בחינוך ילדיהם בבית, לא בבית הספר. [הם]1 צריכים לדאוג שהילדים של [הם]1 לא יכלו זמנם במשחקי מחשב; [הם]1 צריכים לוודא שהילדים מבלים מספיק זמן בהכנת שיעורי בית; [הם]1 צריכים להגיע לאספות הסמסטריאליות בבית הספר. [הורים]2 נוטים להאשים את בתי הספר במגבלות החינוכיות של הילדים של [הם]2. אם [הורים]3 אינם מרוצים מבית ספר, צריכה להיות ל[הם]3 אפשרות לעבור לבית ספר אחר.

אגד אזכורים 1: {הורים, הם, הם, הם, הם}
אגד אזכורים 2: {הורים, הם}
אגד אזכורים 3: {הורים, הם}

כמו כן צירופים שמניים שאינם ספציפיים כמו "תנור לבנים" או "מחשב מקבוק פרו" ייחשבו כביטויים כלליים ובמקרה שיופיעו במשפט "תנור לבנים" ו"טאבון" הם לא יקושרו אחד לשני.

[תנור לבנים], המכונה לעיתים [טאבון] הוא מתקן עשוי אבנים או לבנים שמסיקים בו אש ומיועד בעיקר לחימום, לאפייה ולבישול.

אגד אזכורים 1: {תנור לבנים}
אגד אזכורים 2: {טאבון}

כינויי גוף כלליים: כינויי גוף כלליים, הם כינוי גוף אשר לא מתייחסים לאף ישות ספציפית, אלא מבטאים התייחסות לקבוצה או ישות אבסטרקטית. כינויים כאלה יהיו לא יהיו חלק מאגד אזכורים אחר ויהוו אגד יחידני (singleton).

תהיה [אתה] השינוי ש[אתה] רוצה לראות בעולם. *(מהטמה גנדי)*

תיאורים מאייכים

מרכיב לשוני כלשהו הוא תיאור מאייך אם הוא מצביע על ייחוד כלשהו בביצועה של פעילות נתונה שמתממשת לשונית בפועל ; או הוא מצביע על ייחוד דומה במה שחל בהתרחשות מסוימת , במצב מסוים העשויים גם הם להתממש לשונית בפועל נתון .[3] דוגמא לכך היא מילים כגון "היטב", "כביכול", "מאד": לדוגמא באזכור *ילדה יפה מאד* המילה "*מאד*" מאייכת את שם התואר יפה. אך הייחוד יכול להיות גם באמצעות תיאורים כגון "בשקט": *הילדים הלכו הבייתה בשקט.*

תיאורים מאייכים לא יעמדו בפני עצמם כחלק מאגד אזכורים ולא יקושרו לאזכור אחר אם אינם שמות עצם ומשמשים כתואר הפועל. כמו כן, בדומה לשמות עצם מרוכבים, תיאורים מאייכים אשר אינם שמות עצם יהוו חלק מהאזכור של שם העצם שהוא שורשם.

[ילד קטן מאד] הלך לגן, [הוא] התיישב במעגל ליד הגננת.
אגד: {ילד קטן מאד (אזכור עם תיאור מאייך), הוא}

הילדים הלכו הבייתה ב[שמחה]. [השמחה] הייתה רבה כשהופתעו לגלות שגם מחר אין בית ספר.

- ה "שמחה" הראשון הוא תיאור מאוייך שמשמש כתואר הפועל ולכן לא יקושר ל"השמחה" למרות שהמשמעות דומה.

ביום ראשון בערב שני תושבי צפון הארץ, אב ובנו, פשטו על [מקשת אבטיחים] השייכת לחקלאי מטירת צבי והעמיסו לטנדר שלהם אבטיחים בשווי כולל של עשרות אלפי שקלים. מתנדבי ארגון "השומר החדש" ולוחמי מג"ב הבחינו ברכב החשוד עמוס ב[אבטיחים] נוסע ליד אחד היישובים בצפון הארץ.

- למרות שנראה כי "מקשת אבטיחים" ו"אבטיחים" עלולים להצביע על אותו אובייקט, הם לא יהיו חלק מאותו אגד אזכורים.

[3] גלילה מורשם הספר: כיצד מתארים: עיוני תחביר ומשמעות בעברית בת-ימינומקום ההוצאה: ירושליפשם ההוצאה: מוסד ביאליק שנת ההוצאה: 2013עמוד: 25

ביטויים מאייכים אשר עומדים בפני עצמם

יוצא דופן לסעיף הקודם שאינו מקשר תיאורים מאייכים הם תיאורים מאייכים של זמנים כגון "*יולי*", "*השנה*" ו"*מחר*".

ב[דו"ח [יולי]1] מצויין כי רווחי החברה עלו ב20% ב[חודש זה]1.
אגד אזכורים: {יולי, חודש זה}

כמו כן, ביטויים מאייכים שהם יישות, יסומנו גם כן[4] באופן נפרד ויוכלו להיות מקושרים לאזכורים אחרים בטקסט.

מזכיר [הקיבוץ החדש] נשא את נאומו לפני הקהל שהורכב ברובו מיוצאי [הקיבוץ].
אגד אזכורים: {הקיבוץ החדש, הקיבוץ}

שמות

שמות לא יפורקו ונתייחס אליהם כמקשה אחת, בתוך זה נכלול תארים כגון "מר", "ד"ר" וכדומה אשר לא יופרדו משם העצם שהוא שורשם. אי הפירוק כולל בתוכו גם **שמות של מקומות**, או **שמות שירים** אשר לא יקושרו לאזכורים אחרים.

[פרופסור דני אורבך] התארח הערב בתוכנית פגוש את העיתונות. [הוא] רואיין בנוגע לספרו החדש בנוגע ליפן.
אגד אזכורים: {פרופסור דני אורבך, הוא}

בשיר [בית המלון בקליפורניה] שרה להקת הנשרים על מלון ב[קליפורניה]. זהו [השיר המפורסם ביותר של הלהקה].
אגד אזכורים: {*בית המלון בקליפורניה*, *השיר המפורסם ביותר של הלהקה*}

- למרות שקליפורניה מופיע כחלק משם השיר, חלק זה לא יקושר אזכור קליפורניה.

אני כבר שנים לקוח ב[בנק ירושלים] מאז שגרתי ב[בירת ישראל].

- למרות שירושלים מופיע כחלק מהאזכור "*בנק ירושלים*" הוא לא יקושר לאזכור "*בירת ישראל*"

ביטויים מכמתים

ביטויים מכמתים כפי שהוגדרו בפרק הגדרת האזכורים לא יתוייגו באותו אגד של שם עצם והביטוי המכמת שלו.

- *[שלושה [אנשים]]x*
- *[הרבה [חיות]] x*
- *[מקטע של [כביש 6]]x*
- *[נחיל [דגי דקר]]x*

[4] מבחן טוב לבדיקה האם אזכור הוא יישות או לא, הוא האם כאשר נתרגמו לאנגלית יהיה עלינו לכתובו עם אות גדולה (Capital letter) או לא

האזכור הארוך יותר (עם הכמת) והאזכור הקצר יותר (ללא הכמת) לא יפנו אחד לשני ולא יהוו כשני אזכורים נפרדים באותו האגד אלא נשתמש בביטוי כולל הכמת עבור הפניות אחרות אליו. באופן ספציפי גם עבור כמת כמו "כל" אנו נפעל באותו האופן.

[כל [הילדים בגן]] הלכו לישון. [הם] קמו לקראת ארוחת הצהריים.
אגד אזכורים: {כל הילדים בגן, הם}

[שלושת [הילדים בגן]] הלכו לישון. [הם] קמו לקראת ארוחת הצהריים.
אגד אזכורים: {שלושת הילדים בגן, הם}

רשימות

כאשר מופיע רשימה אשר קיימת אליה הפנייה במסמך, אנו נסמן את כל פרטי הרשימה כאזכור אחד והאגד יורכב משני פריטים: כל חברי הרשימה וההפנייה אליהם.

את הקורס מבוא למדעי המחשב באוניברסיטה העברית מלמדים השנה [פרופ' אביב זהר ,פרופ' ג׳פרי רוזנשיין וד"ר אריה שלזינגר]x. [הם]x מעבירים את הקורס יחדיו כבר שנים רבות.
אגד אזכורים: {"פרופ' אביב זהר ,פרופ' ג׳פרי רוזנשיין וד"ר אריה שלזינגר", "הם"}

מקרים מיוחדים

ארגונים ותתי ארגונים

במידה ומופיע ארגון או חברה, וחלק מהארגון או חברה במסמך אנו לא נאגדם יחדיו.

[משרד התחבורה]1 הודיעה אתמול שכ[200 מעובדיו]2 ישבתו מחר אחרי שהמשא ומתן בין [הם]2 ל[משרד]1 לא צלחו.
אגד אזכורים 1: {משרד התחבורה, משרד}
אגד אזכורים 2: {200 מעובדיו, הם}

אי התאמה במגדר או מספר

במידה וקיימת אי התאמה של מין או מספר בין שני אזכורים אך ניתן להבין שהם מצביעים על אותו אובייקט בעולם אז אנו עדיין נאגדם יחדיו.

הזמנתי [כוס קפה] בארומה, אך [הוא] עדיין לא מוכן.
אגד אזכורים: {כוס קפה, הוא}

GPE וממשלות

מדינות וממשלותיהן לא יופנו אחת לשנייה. כמו כן, נתייחס למדינות ואזרחיהן בצורה שונה

עם קום המדינה קיוותה [ממשלת [ישראל]] שעצרת האו"ם תקבל את תוצאות מלחמת העצמאות ואת חלוקת העיר בין [ישראל] לירדן, אולם בעצרת האו"ם התגבש רוב ברור להענקת מעמד בין-לאומי לירושלים.
אגד אזכורים: {ישראל, ישראל}

- ממשלת ישראל אינה מסומנת באותו האגד עם ישראל

לאחר מלחמת ששת ימים בה ניצחה [ישראל] [אזרחי [המדינה]] חגגו ברחובות.
אגד אזכורים: {ישראל, המדינה}

שלילה

צירוף שמני שמחובר אליו כמת שלילה לא יסומן יחדיו באותו באגד עם אובייקט אחר, גם אם נראה שמדובר על אותו האובייקט. היות ולמעשה הצירוף השמני עם השלילה מהווה קבוצה ריקה שלא ניתן להצביע עליה בעולם האמיתי.

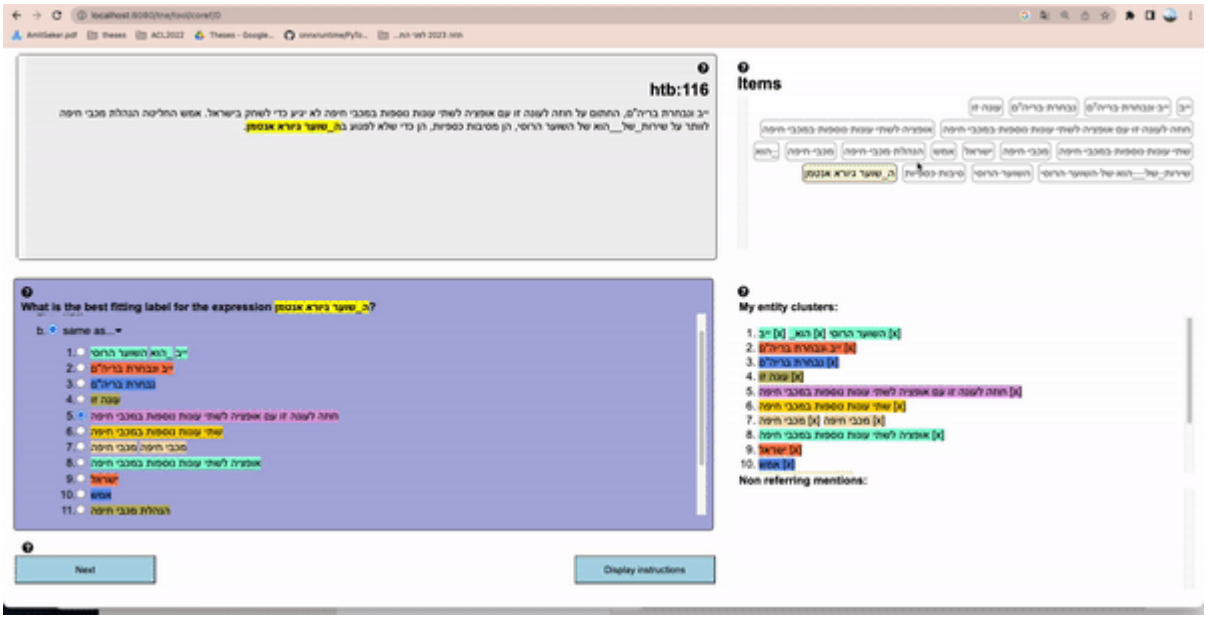
[אף ילד] מבית מהכיתה לא הגיע היום לבית ספר. [כולם] היו חולים.
אגד אזכורים: {כולם}

אין לי [כדורגל]. ממש הייתי רוצה לקנות [אחד].
אגד אזכורים: {אחד}

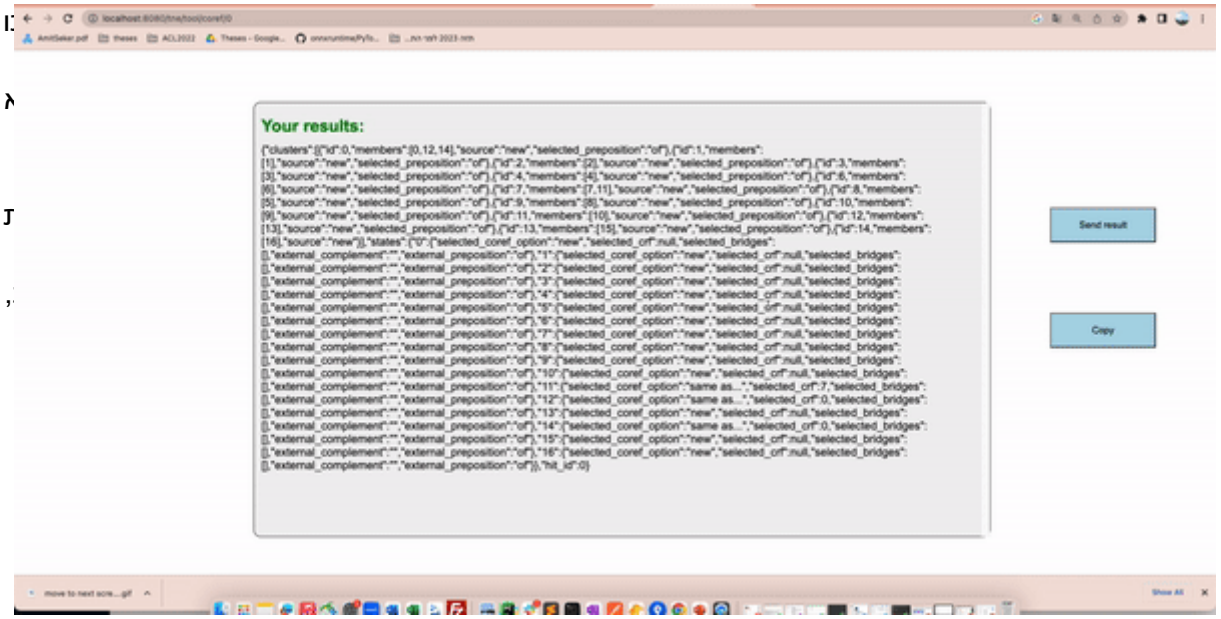
תוכנת התיוג

אנו נשתמש בתוכנת התיוג TNE שפותחה במעבדה בבר אילן. בעת השימוש, יוצג מסמך אחד שלם ורשימת האזכורים (items) שחולצו מהמסמך. בשלב זה תתבקשו לייצר את קבוצות האזכורים השונות, כך שכל קבוצה מכילה בתוכה את כל האכורים לאותה יישות ספציפית.

עבור כל אזכור יוצגו 3 אפשרויות:



1. אגד חדש (new) - נבחר כאשר עדיין לא קיים אגד עבור האזכור הנוכחי
2. צירוף לאגד קיים (same as) - נצרף את האזכור הנוכחי לקבוצה אשר מצביע לאותה ישות כמו הנוכחי.
3. אי הכללה כאזכור (idiomatic) - במידה והקטע המסומן אינו מצביע לאף ישות ספציפית בעולם, אנו נסמנו כidiomaic



יומן שינויים

שינויים עבור גרסא 7

1. התייחסות לתוצאות משחקים או תיאורים שמגיעות בסמוך לשורש בסוגריים
2. שינוי הסדר של מבנים קופולריים להיות אחרי אוגד
3. הוספת דוגמאות לביטוי שלילה.
4. הוספת דוגמאות לאיגודים בהם האזכור הוא חלק מציטוט

שינויים עבור גרסא 6

1. הסבר לגבי ביטויים אידיאומטיים כמו "אין חדש תחת השמש", "עולם כמנהגו נוהג", וביטויים כמו "אף על פי"
2. הסבר לכך שבשמות עצם המתוארים עם פסיק יש לקחת את הביטוי הספציפי ביותר כולל הפסוקית לאחר הפסיק

שינויים עבור גרסא 5

1. התייחסות לצורה בה מסמנים אזכורים חדשים
2. התייחסות לגרשיים וסוגריים
3. תיקון הסבר על סימון אוטומטי של אזכורים
4. הסבר מפורש על מבנים קופולרים שהקופולה מתארת באופן חח"ע את הנושא
5. ניסוח מחדש של מגוון פסקאות ותיקון טעויות סופר וניסוחים.
6. שיפור ההסברים על אזכורי שם תואר ותואר הפועל

שינויים עבור גרסא 4

1. התייחסות לשמות פעולה (שימוש בסמים, ריקודי עם) - כל שמות הפעולה בין אם נחשבים לאירועים ובין אם לאו.

שינויים עבור גרסא 3

1. התייחסות לאמות מידה
2. התייחסות לשמות פעולה (שימוש בסמים, ריקודי עם)
3. שמות עצם שמשמשים כתיאורים (הוא רואה בו מס מעוות)

