עבודת גמר בבלשנות עברית בנושא:

בדיקת היתכנות לאפיון מילים רב משמעיות באמצעות סממנים חיצוניים בטקסט עברי

שם התלמיד: יותם ענבר

תעודת זהות: 213135221

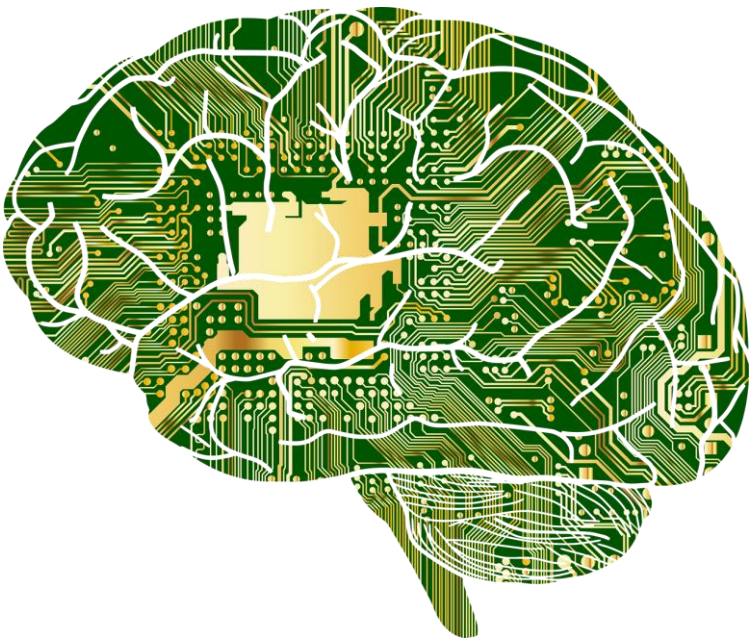
טלפון: 0584262484

שם בית הספר: תיכון עירוני ד' ע"ש אהרון קציר

סמל בית הספר: 540146

טלפון בית הספר: 03-6053665

שם המנחה: מר ברק פז



תוכן עניינים

הקדמה אישית

הרעיון לכתיבת עבודת גמר בבלשנות עלה כשהייתי בסוף כיתה י' ומקצוע הלשון היה אחד מהמקצועות האהובים עלי. כשהתעניינתי לגבי האפשרות להרחיב על המקצוע דרך בית הספר, על אף שגיליתי שלא קיימת אפשרות להיבחן בבחינת בגרות בלשון בהיקף 5 יחידות לימוד, הוצגה בפני האפשרות לכתיבת עבודת גמר במקצוע. שנה שלמה של עבודה מאומצת על ההצעה הביאה אותי להבנה שאני לא יודע לתכנת ושאולי כדאי לחשוב מחדש על הכתיבה של העבודה. בסבלנות והבנה רבות מצד המנחה שלי, ברק, הצלחנו להגיע לנושא חדש שקרוב יותר לנושאים שמעניינים את שנינו. חצי שנה של התחבטות בקוד, מבחנים סטטיסטיים וביקורת מתמדת מצד ברק לאחר מכן, העבודה מוכנה ואני מאוד גאה בה.

ישנם מספר אנשים שלהם אני רוצה להודות מעומק לבי:
הורי, שקראו את העבודה מתחילתה ועד סופה והעירו לי שסוגריים ונקודה-פסיק אינם סימני פיסוק אקדמיים ושמשפטים של שלוש שורות הם לא תקינים;

הגברת לימור שיאון, שליוותה אותי ואת שאר חברי לשכבה בכתיבת העבודות, ותמיד היתה מבינה וששה לעזור בתסבוכות בירוקרטיות;

הגברת עפרה לשם, מורתי ללשון, שהחדירה בי את האהבה לחקר השפה העברית, תמיד הייתה מוכנה לענות לי על שאלות מתחכמות על מקרים יוצאי דופן שאינם בחומר, וסיפקה לי חומרים להעשרה כשסיימתי לקרוא את ספר הלימוד בבית;

וכמובן, המנחה הנפלא שלי, מר ברק פז, שסבל אותי במשך שנה וחצי ושתי הצעות מחקר. הרעיונות הנפלאים שלו לנושאים הם תמיד שילבו עניין רב בנוסף לפתח ליישום בעולם האמיתי. בלי ההצעות, התיקונים, הטלות הספק, התמיכה והידע הנרחב שלו, אני לא יודע מה הייתי עושה.

בלעדיכם העבודה הזו לא הייתה נכתבת.

מבוא

ישנן שפות, כגון עברית וערבית (שבנויות על מערכת כתב עיצורית, "אבג'ד"), להן תווים או אותיות רק לציון עיצורים, והתנועות מסומנות בעזרת סימנים דיאקריטיים שונים או בעזרת אמות קריאה (אותיות המסמנות לעתים עיצורים ולעתים תנועות, כמו האותיות העבריות א,ה,ו',י), אך לרוב הן לא מסומנות בכלל. עובדה זו יוצרת מילים שונות שגם מבוטאות באופן שונה, אך נכתבות ללא ניקוד באופן זהה. על הקורא להחליט בעצמו איזה מן הפירושים של המילה הלא-מנוקדת הוא הנכון באמצעות הקשר. בפרט, על הקורא לזהות את חלקי הדיבר האפשריים של המילה, ולהחליט איזה מהם הוא המתקבל ביותר על הדעת בהינתן שאר המשפט. לתהליך זה קוראים "הפגת עמימות מורפולוגית", ובו הקורא עשוי להידרש להפגת עמימות מסוג אחר, אם ייתקל במספר פירושים שונים אך זהים מבחינה מורפולוגית.

את תהליך הפגת העמימות המורפולוגית ניתן לבצע באמצעות מחשב, כשיש להזין את המחשב ברשימה של כל חלקי דיבר של כל מילה במשפט ונתונים סטטיסטיים עליהם והוא יפלוט את רצף חלקי הדיבר הסביר ביותר של כל מילה. בתהליך זה אי יעילות מסויימת, שכן יש לבדוק בו כל צירוף אפשרי של חלקי דיבר, אך הבדיקה המתמטית נעשית על צמדי מילים כל פעם (והרי את המילים החד-משמעיות אין צורך לבחון כלל). עדיף היה אילו היה ניתן לזהות ראשית באמצעים זולים חישובית את המילים הרב-משמעיות ולהפעיל את האלגוריתם המתמטי רק עליהן. בעבודה זו תיבחן אפשרות דומה. כפי שיפורט בגוף העבודה, ננסה לתת אפיון חיצוני למילים רב משמעיות בהתבסס על מספר קטעי טקסט בעברית. כמו כן, יוצג גם אלגוריתם של הפגת עמימות ממוחשב שמסתמך על היכולת של מחשב לאתר מילים רב משמעיות מבעוד מועד. מעבר לכך, לפי בר-חיים, סמעאן ו-וינטר (2008), הצורך לא להידרש לבדוק כל צירוף אפשרי של חלקי דיבר גדל דווקא בעברית, שהיא שפה עמוסה מורפולוגית , כלומר, למילים רבות בה יש מספר רב של חלקי דיבר שאפשר לשייך להן.

פרק 1 - מורפולוגיה מודרנית

מורפולוגיה היא תחום בבלשנות העוסק ברכיבים בעלי משמעות בתוך המילה, באופן בו הם מצטרפים למילים שונות ובמשמעות אותה הם נותנים. היחידה הבסיסית של המורפולוגיה מכונה "צורן" ("מורפמה בלעז). צורנים זהים מקנים למילים שונות משמעויות זהות - למשל, הצורן 'ה-' בהצטרפו לשם עצם תמיד יציין יידוע, בין אם במילה "הבתים" ובין אם במילה "המעגל".

במורפולוגיה המודרנית מספר תחומי מחקר. עבודה זו תעסוק בשניים מהם:

סינתזה, העוסקת בדרך בה מילים נגזרות מצורנים. הגזירה בעברית נעשית על צורני בסיס המכונים גזעים או שורשים באמצעות מוספיות - תחיליות (בתחילת מילה, והילקוט), סופיות (בסוף מילה, 'ביתי', (תוכיות (בתוך מילה, 'לימד') ומוספיות מסגרת (מסביב למילה, 'תאכלו'). התחיליות והסופיות מייצרות נטייה המכונה "גזירה קווית", מכיוון שהיא משאירה את הגזע שלם, ואילו התוכיות ומוספיות המסגרת יוצרות "גזירה משורגת" שכן הצורנים בה משולבים זה בזה (הגזע לא נשאר שלם). כמו כן, המוספיות בעברית לרוב יתווספו על בסיסים המכונים שורשים של שלושה עיצורים. קיימים גם חוקים להשתנות המוספיות במקרים בהם עיצורים אלה כוללים אותיות כגון ו', י' ו-א'.

אנליזה ,בה נחקרת ההפרדה של מילים לצורנים שמרכיבים אותן, ובזיהוי חלק הדיבר של מילים בתוך משפט[1]. חלק הדיבר של מילה הוא דרך לסווג אותה לפי מאפייניה המורפולוגים. בעברית קיימת חלוקה לשמות עצם, פעלים, שמות תואר, מילות יחס ותוארי הפועל. רבים מהחוקים התחביריים והמורפולוגיים בעברית מנוסחים לפי חלקי דיבר, כשלדוגמה, מילת יחס בעברית תבוא רק לפני שם עצם.

ישנן מילים בעברית, כגון המילה 'של', שכשהן נכתבות בלא ניקוד, ניתן לייחס להן מספר חלקי דיבר אפשריים - 'של', למשל, תתפרש כשם עצם ("של כחול על סוודר תכלת"), מילת יחס ("הבית של פיסטוק") וכפועל בציווי ("של נעליך מעל רגליך"). זו אינה בעיה כאשר נעשית אנליזה "ידנית" (בידי אדם): בדוגמה הראשונה, נבחין שלאחר המילה 'של' בא שם תואר בזכר. אפשרות זו תואמת רק שם עצם בזכר, ולכן זו המשמעות הסבירה למילה 'של'. עם זאת, נוצרת בעיה כאשר נעשה ניסיון להפוך את תהליך האנליזה לממוחשב, שכן למחשב אין היכולת לבחור את הניתוח הנכון ללא אלגוריתם מתאים. יצויין שגם בני אדם משתמשים באלגוריתמים, אך הם מורכבים מכדי לתרגם אותם ישירות למחשב. לשם השגת הניתוח הנכון המחשב משתמש בגרסה ראשונית של הבנה מהקשר: בתהליך זה , המכונה "הפגת עמימות מורפולוגית", המחשב משתמש במידע על המילים שמסביב לכל מילה בכדי לייצר משפט שמורכב מהניתוחים הכי סבירים של כל מילה. בפרט, המחשב משתמש בהסתברויות של הופעת צמדי ניתוחים שונים, שמוזנות לו מתוך ניתוח ידני של קורפוס ארוך ומייצג. להלן האלגוריתם הנ"ל כפי שמחשבים משתמשים בו כיום, מופעל על המשפט "ילדה אכלה את התפוח". האלגוריתם בדוגמה זו מופעל על נתונים מומצאים, שנלקחו מתוך ספר הלימוד "חקר השפה - יסודות ויישומים" (ראה ביבליוגרפיה):

ראשית, מופקת טבלה של כל הניתוחים האפשריים של כל מילה במשפט, בצירוף של ההסתברות שכל ניתוח יופיע באופן עצמאי. למשל, המילה 'את' כמעט לא מופיעה במשמעות שם עצם.

[1] שני תחומים אלו קשורים באופן הדוק, שכן בכדי לזהות את חלק הדיבר של מילה יש ראשית "להפשיט" אותה מהמוספיות שלה (המילה 'שבתה' עשויה להתפרש כפועל, אך גם כ-'ש' השיעבוד+שם עצם+כינוי קניין, כדוגמה).

טבלה 1 : ההסתברות שמילה תופיע בחלק דיבר מסויים

חלק דיבר / מילה	תואר	מיידע	מילת יחס	שם עצם	כינוי גוף	פועל
ילדה	0	0	0	80	0	20
אכלה	0	0	0	0	0	100
את	0	0	83	3	14	0
ה	0	100	0	0	0	0
תפוח	22	0	0	78	0	0

אחרי כן מופקת טבלה שניה שמסכמת את ההסתברות שכל צמד חלקי דיבר יופיע ברצף. כדוגמה, הצירוף מילת יחס-שם עצם נפוץ בהרבה מהצירוף פועל-כינוי גוף.

טבלה 2 : ההסתברות שכל צמד חלקי דיבר יופיע ברצף

שני בצמד / ראשון בצמד	תואר	מיידע	מילת יחס	שם עצם	כינוי גוף	פועל
תואר	10	17	0	0	0	73
מיידע	36	0	0	64	0	0
מילת יחס	0	47	0	53	0	0
שם עצם	15	13	52	11	0	39
כינוי גוף	14	0	22	24	0	40
פועל	0	8	48	40	4	0

לבסוף, בתוספת לטבלה של כל צירופי הניתוחים האפשריים ובשיטות חישוביות שונות, ניתן להצליב את שתי הטבלות הללו בכדי לספק "ציון" לכל צירוף ניתוחים.

טבלה 3 : הציון של כל שילוב חלקי דיבר אפשרי

ילדה	אכלה	את	ה	תפוח	ציון המשפט
פועל	פועל	שם עצם	מיידע	שם עצם	0
פועל	פועל	שם עצם	מיידע	שם תואר	0
פועל	פועל	מילת יחס	מיידע	שם עצם	0
פועל	פועל	מילת יחס	מיידע	שם תואר	0
פועל	פועל	כינוי גוף	מיידע	שם עצם	0
פועל	פועל	כינוי גוף	מיידע	שם תואר	0
שם עצם	פועל	שם עצם	מיידע	שם עצם	2.43
שם עצם	פועל	שם עצם	מיידע	שם תואר	0.236
שם עצם	פועל	מילת יחס	מיידע	שם עצם	292
שם עצם	פועל	מילת יחס	מיידע	שם תואר	41.5
שם עצם	פועל	כינוי גוף	מיידע	שם עצם	0
שם עצם	פועל	כינוי גוף	מיידע	שם תואר	0

בדוגמה זו הציון שניתן לכל משפט הוא מכפלת ההסתברוית עבור כל מילה כפול מכפלת ההסתברויות עבור כל צמד מילים[2]. באופן זה, משפטים המכילים צירופים שלא קיימים בעברית (לדוגמה, הצירוף פועל-פועל) מקבלים ציון אפס. כעת, הניתוח "הנכון" של מילה מוגדר כניתוח שלה במשפט עם הציון הכי גבוה - זהו משפט 9 בדוגמה הגולמית לעיל. לכן, במשפט לעיל, מחשב יסיק שהניתוח של המילה 'את' הוא "מילת יחס" וכיו"ב.

כפי שניתן היה לראות בדוגמה הקודמת, לחלק גדול מהמילים בעברית מספר רב של חלקי דיבר שניתן לשייך להם כשהללו נכתבות ללא ניקוד - מניתוח של קטע קצר, חלק זה לעתים עולה על מחצית. למילים אלה השפעה רבה על קושי ביצוע הפגת עמימות מורפולוגית - מילה עם X ניתוחים אפשריים תכפיל את מספר המשפטים שיש לבדוק פי X (לעומת מילה עם ניתוח אפשרי אחד).

עם זאת, הכפלה זו קורית מכיוון שאנו בודקים את כל המילים הרב משמעיות "בבת אחת" - אנחנו אמנם בודקים את המילים שתיים-שתיים, אך נותנים ציונים למשפטים שלמים (בדוגמה לעיל, את הצירוף 'ילדה' (שם עצם) 'אכלה' (פועל) אנחנו בודקים יותר מפעם אחת - שש, למעשה). בדיקת כל שילובי הניתוחים האפשריים גורמת לכך שסך המשפטים שיש לבדוק הוא מכפלת הניתוחים. במשפט הדוגמה, ישנן שלוש מילים רב משמעיות ('ילדה', 'את', 'תפוח') להן 2,3 ו-2 ניתוחים אפשריים, בהתאמה. בכדי לבחון כל צירוף של חילקי דיבר, יש לבדוק 12=3*2*2 משפטים.

[2] כמוזכר, האלגוריתם לחישוב הציון עבור כל משפט הוא מעט מתוחכם ממכפלה פשוטה של כל הציונים במשפט. אלגוריתם זה מוסבר במאמרם של בר חיים, סימעאן ו-וינטר (2008)

אם נוכל, כפי שיוצג, לזהות את המילים הרב משמעיות מבעוד מועד, נוכל לבדוק כל אחת מהן בנפרד. עבור כל אחת מהן, נצליב את שתי ההסתברויות שלהלן: ההסתברות של הופעת המילה בחלק דיבר מסויים, וההסתברות שחלק דיבר זה יופיע ברצף עם חלק הדיבר של המילים שלפני ואחרי המילה. כך נצטרך לבצע את ההצלבה פעם אחת עבור כל מילה רב משמעית. למעשה נצטרך לבדוק מספר משפטים השווה לסכום מספרי הניתוחים, כשאנו מחשיבים את מספר הניתוחים של מילה חד-משמעית ל-0. סכום זה של מספרי הניתוחים לעולם לא יהיה גדול יותר מהמכפלה שלהם[3], ולרוב יהיה אף נמוך יותר: במשפט להלן, יהיה עלינו לבדוק 3+2+2=7 משפטים בלבד, כמעט מחצית מ-12 המשפטים שנוצרו כשבדקנו כל צירוף חלקי דיבר אפשרי.

האלגוריתם שיאפשר זאת עשוי להיראות כך:

1. נעבור על כל מילה במשפט ונבחן אם היא רב משמעית על סמך הקריטריונים החיצוניים.
2. עבור כל מילה רב משמעית, נבדוק מה הם כל הניתוחים האפשריים שלה ובאילו הסתברויות הם באים.
3. עבור כל מילה רב משמעית ושתי המילים שמצדיה, נבדוק מה הם הסיכויים שהניתוחים שלהם יבואו ברצף.
4. עבור כל צמד כזה, נצליב את הנתונים משלב 2 ו-3.
5. עבור כל מילה רב משמעית, הניתוח שקיבל את הציון הגבוה ביותר ייבחר כניתוח הנכון.

אנו מקווים, כמובן, שאלגוריתם זה יעיל יותר מהאלגוריתם המקובל. יעילות זו תתקבל רק אם מחיר זיהוי המילים הרב משמעיות בשלב 1 הוא זול מספיק כדי שהוא לא יתקזז עם עלות ההוזלה שביכולת לנתח מילה-מילה, כשקיים גם הסיכוי שתנאי זה יתקיים במשפטים מסוימים ולא באחרים. יצוין שעברית, בדומה לשפות כמו טורקית וערבית, היא שפה עמוסת מורפמות (שפות אלה מכונות "סינתטיות"), ועל כן ההוזלה המדוברת צפויה להיות משמעותית.

[3] קיימת לכך הוכחה מתמטית, אך היא לא תובא פה. הוכחה דומה מובאת באתר mathoverflow (https://mathoverflow.net/questions/16684/when-is-the-product-of-a-set-of-numbers-greater-than-the-sum-of-them).

פרק 2 - אפיון חיצוני של מילים רב משמעיות

ניתן לאפיין מילים בעזרת מגוון קריטריונים - למשל, שימושן בתחומי החיים (לדוגמה: "מילה המתארת מכשיר חשמלי"). אפיונים מעין אלה עשויים להועיל רבות במציאת מילים רב משמעיות אך ישנו קושי חישובי רב בלימוד מחשב להשתמש בקריטריונים אלה. קושי זה מחטיא את המטרה של האפיון השטחי, שהיא כאמור הייעול החישובי של תהליך הפגת העמימות המורפולוגית.

אם כן, ננסה להשתמש רק בקריטריונים שנחשיב לאלמנטריים או "חיצוניים". הקריטריונים הם להלן: אורך המילה (2,3,4,5 ו-6 אותיות), הימצאות תחילית (ב' ,כ' ,ל' ,מ', ה', ו', ש'), מיקומה של המילה במשפט (תחילת משפט, סוף משפט), הימצאות סיומת ('ים', 'ות', י'). קריטריונים אלה נבחרו באופן מעט שרירותי - למשל, לא בחנו את כל התחיליות האפשריות בעברית, אלא רק את אלה שיש להן משמעות כצורנים (תחיליות בכל"ם המציינות מילות יחס, ה' היידוע, ו' החיבור ו-ש' השיעבוד). באופן דומה נבחרו הסיומות והמיקומים. לאלה נוספו הקריטריונים "המילה פותחת ב-ע'" ו"המילה מסתיימת ב-א'". לאותיות א' ו-ע' אין תפקיד בתור אותיות שימוש, ולכן לא מדוייק לכנותן "תחיליות" ו"סופיות". עם זאת, קריטריונים אלה בלטו בציונן לאחר בחינת כל התחיליות והסופיות האפשריות בעברית, וגם הן נוספו למכלול הקריטריונים שנבחנו. כפי שיוסבר בהמשך, בולטות זו נובעת מהחריגות של הסתברות הופעת מילים מסויימות שעומדות בקריטריונים אלה (המילה הרב-משמעית "על", למשל, היא המילה השלישית בשכיחותה בקורפוס). במחקר עתידי יהיה צורך לבצע נרמול של המילים, כך שהטיב של הקריטריונים לא יושפע מחריגות כאלה.

בחירה מוגבלת זו תאפשר לנו להעלות השערות מנומקות בנוגע להשפעה שנצפה שתהיה לקריטריונים. תוך כדי כך, נתחשב בתדירות הופעת המילים לא בתפקיד אות שימוש (באופן "טבעי", כמו האות ה' במילה "הוא"), ובתפקידן כאות שימוש (באופן "לא טבעי", כמו האות ה' במילה "הבית"). יצויין גם שבעתיד, אם ייתאפשר יישום של אפיון זה באמצעות רשתות נוירונים מלאכותיות, הרשת תוכל למצוא את הקריטריונים בעצמה, ואף להגיע לכאלה שלא נחשבים "פשוטים" (כגון "תחילית מ' ושני מקרים של הימצאות י' בתוך המילה"). היא גם תוכל לייצר משקל שונה לכל קריטריון באופן שמאפשר, כך ניתן לקוות, לאפיון מהימן הרבה יותר משיושג בעבודה זו.

תחיליות:

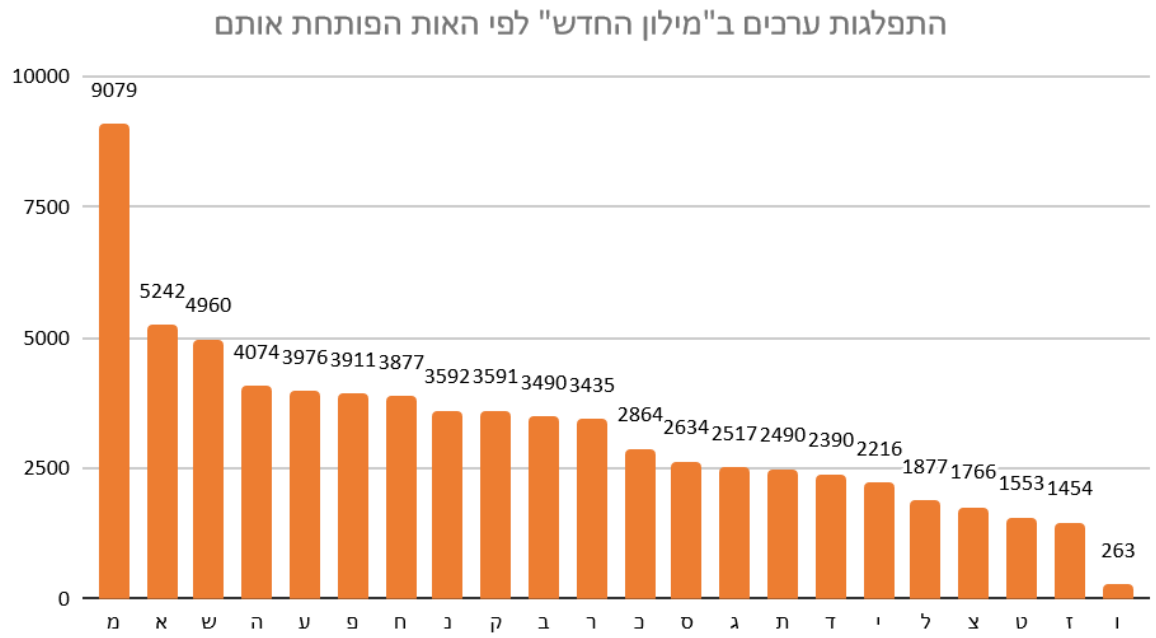
ב' - מאותיות היחס. אותיות אלה מחליפות מילות יחס שלמות (בתוך, כמו, אל, מן...), ועל כן הן נוטות להיות מאוד נפוצות. כמו כן, כאשר אותיות אלה מופיעות באופן "לא טבעי" (כאות שימוש) עשוי לבוא אחריהן שם עצם בלבד. על כן, נצפה שככל שהמילה תהיה נפוצה יותר באופן זה, כך הסיכוי שלאחריה תגיע מילה רב משמעית יקטן, ולהפך - ככל שהמילה תהיה נפוצה יותר באופן "טבעי", כך יגדל הסיכוי שלאחריה תגיע מילה חד משמעית. בכדי לקבל אומדן של שכיחויות התחיליות ה"טבעיות", נבחן את התפלגות הערכים המתחילים בתחילית זו ב"מילון החדש" מתוך 71251 המילים שבמילון (סיכום התפלגות זו מובא בגרף העמודות להלן). האות ב', כפי שניתן לראות בגרף, היא לא נפוצה במיוחד ולא נדירה במיוחד (היא פותחת 3490 ערכים), ועל כן נצפה שהקריטריון "המילה פותחת באות ב'" לא יראה באופן מובהק על מילה רב משמעית ולא על מילה חד משמעית.

כ' - מאותיות היחס. האות כ' מופיעה באופן "טבעי" בשכיחות נמוכה מזו של האות ב' (היא פותחת 2864 ערכים), ועל כן נצפה שהקריטריון "המילה פותחת באות כ'" ירמז על הימצאות מילה חד משמעית, במובהקות גבוהה מעט מזו של הימצאות תחילית ב'.

ל' - מאותיות היחס. לפי הגרף להלן, ישנו מספר מועט מאוד של ערכים במילון הפותחים בה. על כן, שהסיכוי שמילה תפתח דווקא בל' היחס, כלומר תהיה בוודאות שם עצם, גבוה, והקריטריון "המילה פותחת באות

ל״י עשוי להורות על הימצאות מילה חד משמעית. עם זאת, האות ל׳ פותחת את כל שמות הפועל בעברית, צירוף שיוצר לעתים רבות רב-משמעויות (ללמוד: ל-למוד), על כן הקריטריון עשוי להיות לא מובהק מאוד או אפילו לרמז על מילה רב משמעית, כתלוי בשכיחות שמות הפועל בקורפוס.

מ׳ - מאותיות היחס. בגרף להלן בולטת שכיחותה של האות מ׳ שכן היא האות הפותחת השכיחה ביותר במילון, עם 9079 ערכים שפותחים בה, ויצויין שמספר זה יהיה אף גבוה יותר אם נחשיב את צורות הבינוני שלא מופיעות במילון כערכים נפרדים. על כן, ניתן לשער שהקריטריון ״המילה פותחת באות מ״ יראה באופן מובהק למדי על נוכחות מילה רב משמעית.



ה׳ - מילית היידוע. ה׳ בתפקיד מיידעת היא נפוצה מאוד, ומכיוון שרק שמות עצם ותארים עשויים להיות מיודעים, ה׳ ״לא טבעית״ תורה על מילה חד-משמעית. אך כפי שניתן לראות בגרף להלן, ה׳ טבעית היא נפוצה מאוד (וכוללת בין השאר פעלים בבניין הפעיל, הופעל והתפעל, וגם שמות פעולה רבים). על כן, ניתן לשער שתחילית ה׳ היידוע לא תראה באופן מובהק לאף כיוון - לא שהמילה חד משמעית ולא שהמילה היא רב משמעית.

ש׳ - מילית השיעבוד. לאחר האות ש׳ השיעבוד מתחיל למעשה משפט חדש (פסוקית), ועל כן נצפה שההשפעה של הקריטריון ״המילה פותחת באות ש״ תהיה דומה לזו של הקריטריון ״המילה נמצאת בתחילת משפט״. מסקירה קצרה של קטע טקסט, עושה רושם שרוב לא מבוטל (למעלה מ-50%) מהמילים בתחילת משפט הן רב-משמעיות. נוכל להסיק מכך שלאחר ש׳ יבואו לרוב מילים רב משמעיות גם כן. עם זאת, מילים הפותחות בש׳ ״טבעית״ הן נפוצות מאוד (היא התחילית הטבעית השלישית בשכיחותה, עם 4960 הופעות), לכן הקריטריון לא יהיה מובהק כמו הקריטריון ״המילה נמצאת בתחילת משפט״.

ו׳ - מילית החיבור. היא עשויה לסמן פתיחה של איבר במשפט מאוחה או להורות על חלק כולל. במקרה הראשון שוב נפתח למעשה משפט חדש, ועל כן גורל המילה שלאחר האות ו׳ יהיה כזה שלאחר האות ש׳. זאת למעט המקרים שבהם האיבר לא נפתח בו׳ ממש אלא במילת חיבור עם ו׳ לפניה כגון ״ולכן״ - המקרה הזה מעניין יותר, מכיוון שכמעט כל חלק דיבר בשפה העברית יכול להופיע כחלק תחבירי כולל, בין אם פעלים (כנשואים כוללים), שמות עצם (כנושאים וכמושאים) או שמות תואר (כלוואים). אם כן, נצפה שהקריטריון ״המילה פותחת באות ו״ ישפיע מעט על הדיוק של המנתח. אם ההשפעה שלו תהיה לטובה או לרעה, זאת ניתן לקבוע בעזרת ההתפלגות של ו׳ בתור פותחת מילת חיבור ובתור פותחת חלק כולל - אם

הראשון נפוץ יותר, כלומר רוב ההופעות של האות ו׳ הן לפני מילות חיבור, סימן שלאחר ו׳ יבואו בעיקר מילות יחס מובהקות והיא תסמן את פתיחתה של מילה חד משמעית. אם השני נפוץ יותר, כלומר ו׳ מופיעה בעיקר לפני האיבר האחרון בחלק כולל, אז אחרי האות ו׳ עשויה לבוא כל מילה וכמעט לא תהיה לקריטריון השפעה. כמו כן, האות ו׳ היא התחילית הטבעית הנדירה ביותר בעברית (רק 263 מילים ב״מילון החדש״ פתחו בה), לכן היא כנראה לא תרמז על מילה רב משמעית - כאמור, את המידה בה היא תרמז על מילה חדמשמעית לא ניתן לשער מראש.

ע׳ - האות ע׳ אינה אות שימוש. עם זאת, היא האות הפותחת של מילת היחס ״על״. מילה זו היא רב משמעית (super-, on), והיא מהווה 13 מתוך 17 המילים הפותחות באות ע׳ (76.5%) בקורפוס עליו נבחנו הקריטריונים. על כן, לכאורה קריטריון זה יהיה מוצלח מאוד בניבוי מילים רב משמעיות - למעשה, קריטריון זה פשוט יהיה מוצלח בניבוי המילה ״על״, שהיא רב משמעית. עובדה זו תודגש בהמשך, אך הקריטריונים רק מנבאים מילים רב משמעיות ולא מאפיינים אותן - מילה רב משמעית לא מתחילה לרוב בע׳ (כלומר, כפי שראינו, היא אינה לרוב המילה ״על״), אך מילה שמתחילה בע׳ (כלומר, בעיקר המילה על) היא לרוב רב-משמעית.

סיומות:

ים/ות - סיומות הרבים בעברית מסמנות ריבוי בשני מקומות: בשמות עצם (כלב > כלבים) ובצורת בינוני (שומרת > שומרות). האותיות ו׳ והן י׳ עשויו באמצע מילה גם כשהן חלק מתבנית וגם כחלק משורש (באופן ״טבעי״), ועל כן הסיומות הללו יתפקדו לעתים בתור מציינות רבים ולעתים בתור חלק משם. על כן, הקריטריונים ״סיומת ׳ות״ ו-״סיומת ׳ים״ יראו חוסר מובהקות לשני הכיוונים, כלומר לא יראו על הימצאות מילה רב משמעיות ולא על הימצאות מילה חד משמעית. כמו כן, מכיוון שגופי הרבים בעברית משמשים גם כגופים סתמיים, הקריטריון ״ים״ עשוי להופיע רבות יותר בתור בינוני בגוף סתמי ועל כן להורות על מילה רב משמעית. על כן, הקריטריון ״סיומת -ים״ עשוי, בניגוד לקריטריון ״סיומת -ות״ להיות להראות באופן לא מובהק על הימצאות מילה רב משמעית.

י׳ - סיומת המציינת שייכות לגוף ראשון יחיד (״ילקוטי״) וגם הפיכה לשם תואר (״סביבתי״). מילים רבות בעברית עשויות להופיע עם שני סוגי הסיומות - למשל, ״ביתי״ במשמעות ״הבית שלי״ (שם עצם + כינוי קניין) וגם ״באופן שמזכיר בית/קשור לבית״ (שם תואר). מילים מסוג זה הן רב משמעיות תמיד, ועל כן נצפה שנוכחות של סיומת י׳ תצביע על מילה רב משמעית.

א׳ - לאות א׳ אין תפקיד כאות שימוש בעברית. כמו כן, סיומת א׳ בעברית הן נדירות - בקורפוס המצומצם עליו נערכה העבודה (504 מילים) נמצאו רק 15 מילים שנגמרו באות א׳ (בניגוד ל-29 הופעות של סיומת י׳). אם כן, הקריטריון של סיומת א׳ יתנהג כמו הקריטריון של תחילית ע׳: הבולטות שלו נובעת מזה שרוב המילים שמסתיימות באות א׳ הן המילה לא על נטיותיה וכינויי הגוף הוא והיא (בנוסף לשני מקרים של הפועל הנטוי יצא). מכיוון שמילים אלה הן חד משמעיות, הקריטריון ״סיומת א׳״ יצביע על נוכחות מילה חד משמעית.

מיקום במשפט:

תחילת משפט - המחקר בתחום התפלגות של מילים בתחילת משפט הוא לא נרחב. עם זאת, ידוע כי סדר החלקים המילים הקנוני בעברית המודרנית הוא נושא-נשוא-מושא[4]. עם זאת, משפטים שחורגים מסדר המילים הזה הם נפוצים ביותר, לפחות בשפה הכתובה, עד כדי כך שכפי שהוזכר קודם, נדמה שמילים רב

[4] זאת על אף שסדר המילים בעברית המקראית הוא נשוא-נושא-מושא (״וידבר ה׳ אל משה...״); למעשה, חוקים דקדוקיים מסויימים בעברית המודרנית, כגון העובדה שפועלי עזר קודמים לפעלים, עדיין תקפים על אף שהם מאפיינים שפות נשוא-נושא-מושא.

משמעיות נוטות להימצא בתחילת משפטים. כאמור, המחקר בתחום זה הוא לא רב, ועל כן בין אם הקריטריון יתברר כמועיל, כגורע או כלא משפיע, יהווה הדבר חידוש.
סוף משפט - אין מאמרים רבים בנוגע למילים בסוף משפט. עובדה זו היא מוצדקת, שכן מבחינה של קטע טקסט קצר נדמה שלא קיים מתאם בין הימצאות של מילה בסוף משפט לבין היותה רב משמעית או להפך. נצפה, אם כן, שההשפעה של קריטריון זה תהיה זניחה.

עתה נוכל לבחון את הקריטריונים. בעזרת קטע הקוד שבנספח 2, נשווה את הקביעה של המחשב האם המילה היא רב משמעית עם הקביעה ה"אמיתית", שהושגה באמצעות כלי הניתוח המורפולוגי של מיל"ה[5] (ולאחר שהוא הפסיק לעבוד, של מורפיקס). המחשב מחזיר "ציון" שמהווה את אחוז הפעמים בהם הקריטריונים תאמו את הניתוח הממשי. ניתוח זה הוא כזכור בינארי - מילה יכולה להיות או חד משמעית או רב משמעית. השוואה זו נעשתה על פני 28 משפטים שנלקחו מתוך שלוש כתבות ("ריח של ים", ד"ר איתן אקסנברג [2019]; "איך הכחדנו את הממותות... שוב", אור אליאסון [2019]; "סרטון הפלסטיק בחסה", מעריב אונליין [2018]), והיא כוללת 504 מילים.

ראשית נבחן כל קריטריון בנפרד. "נכבה" את שאר הקריטריונים מלבד זה שאנו מעוניינים לבחון - באופן מעשי הכיבוי יושג באמצעות הוספת הסימן '#' שהופך את השורות המסומנות בו לתגובה, כך שהמחשב מתעלם מהם. לאחר מכן נריץ את הקוד - הוא יפלוט ציון מספרי, בין 0 ל-100 שמייצג, באחוזים, את מידת הדיוק של הקריטריון; המחשב מניח שכל מילה שעומדת בקריטריון היא רב-משמעית, והציון של הקריטריון הוא היחס בין מספר הפעמים שהנחה זו התבררה כנכונה לבין מספר המילים בקורפוס. הציון שהשיג כל קריטריון מוצג בטבלה להלן, כמו גם בצמצום בגרף מתחתיה.

שנית, נמצא את השילוב האופטימלי של קריטריונים. נתחיל מלבדוק איזה קריטריון מקבל את הציון הטוב ביותר, ונשאיר אותו "דולק". לאחר מכן , "נדליק" (נסיר את תו '#' מתחילת שתי השורות של כל קריטריון שמורה על הפיכה לתגובה) כל פעם קריטריון אחר. נריץ את הקוד ונשווה את הציון (באחוזים) החדש לזה שללא הקריטריון שהוסף - אם הציון החדש גבוה מזה שהושג קודם לכן, נשאיר אותו דולק. אם לא - נכבה אותו. כך נעבור על כל הקריטריונים, ונקבל את השילוב שלהם שייתן את הציון הכי גבוה - האפיון החיצוני המייצג ביותר של המילים הרב-משמעיות. ישנה גם אפשרות לבחון כל קריטריון פעמיים: פעם אחת כפי שהוא מוצג בנספח 2, ופעם אחרת כשהשורה 'CritFilled += 1', שתפקידה להורות למחשב שהקריטריון נענה תשתנה ל 'CritFilled -= 1'. באופן זה נוכל גם לאפיין מילים באופן שלילי - אם, כאשר הודלק הקריטריון, הושג ציון גבוה יותר במקרה השני ('CritFilled -= 1'), סימן שדווקא מילים שלא עומדות בקריטריון הזה נוטות להיות רב משמעיות (למשל, מילים שלא מתחילות באות ב').

נדגים כיצד תיעשה הבחינה של הקריטריון "אורך המילה הוא 5" על משפט לדוגמה - "בערבות סיביר פורחת בשנים האחרונות תעשייה יוצאת דופן באופייה" (זהו משפט 26 בקורפוס שבנספח 2):

1. נריץ את השגרה "WrdPerf" על כל מילה, שתחזיר "אמת" אם המילה היא בת 5 אותיות ו"שקר" אם לא. לדוגמה, עבור המילה "סיביר" השגרה תחזיר "אמת" ועבור "האחרונות" היא תחזיר "שקר".
2. נכניס את ההחזר של השגרה "WrdPerf" על כל מילה לרשימה ("List") באמצעות השגרה "SentPerf". נקבל את הרשימה להלן (א = "אמת", ש = שקר): [ש,א,א,א,ש,ש,א,ש,ש]. זאת כדי לבחון ביעילות רבה יותר את קריטריוני המיקום במשפט - אילו היינו בודקים את הקורפוס כולו, היה עלינו לצרף לכל מילה נתון של מיקומה במשפט (ובכך לעכב משמעותית את ייצור הקורפוס).

[5] מיל"ה, מרכז הידע לעיבוד השפה העברית, הינו כלי המספק מגוון קורפוסים וכלים לעיבוד חישובי של השפה העברית.

3. נשווה את הרשימה שהתקבלה לרשימת האמיתות והשקרים הנכונה, כפי שהתקבלה בעזרת כלי הניתוח של מורפיקס ([א,ש,א,ש,א,ש,א,ש,ש]), ונחלק את מספר ההתאמות המוצלחות במספר המילים במשפט - זהו למעשה אחוז ההצלחות או ה"ציון". ישנן 5 הצלחות (במילה השלישית, השישית, השביעית, השמינית והתשיעית) ו-9 מילים במשפט, על כן הקריטריון יקבל ציון של 5/9*100% = 55.55% (עם הסייג הברור שמשפט זה הוא לא בהכרח מייצג).
4. נבצע את שלושת השלבים הראשונים על כל 28 המשפטים בקורפוס. הציון הכולל של הקריטריון הוא הממוצע של הציונים שלו (הפעלות השגרה SentPerf) עבור על משפט. עבור הקריטריון "אורך המילה הוא 5", ציון זה הוא 45.66, כלומר 45.66 אחוז מן המילים שאורכן 5 אותיות הן אכן רב משמעיות.

להלן הציונים שקיבלו הקריטריונים, כפי שהוסבר לעיל:

הקריטריון (בנפרד)	**הציון (%)**
כאשר המחשב מניח שכל מילה היא חד-משמעית (ללא קריטריונים; אחוז המילים החד-משמעיות)	50.64
אחוז המילים הרב משמעיות (100 פחות אחוז המילים החד-משמעיות)	49.36
תחיליות	
מ'	52.60
ש'	52.89
ה'	48.25
ו'	49.22
כ'	50.98
ל'	48.94
ב'	46.01
ע'	52.49
סיומות	
'ים'	47.28
'ות'	50.03
'י'	52.29

א'	48.53
מיקום במשפט	
תחילת משפט	53.23
סוף משפט	51.21
אורך	
2 אותיות	54.16
3 אותיות	55.24
4 אותיות	56.35
5 אותיות	45.66
6 אותיות	44.28

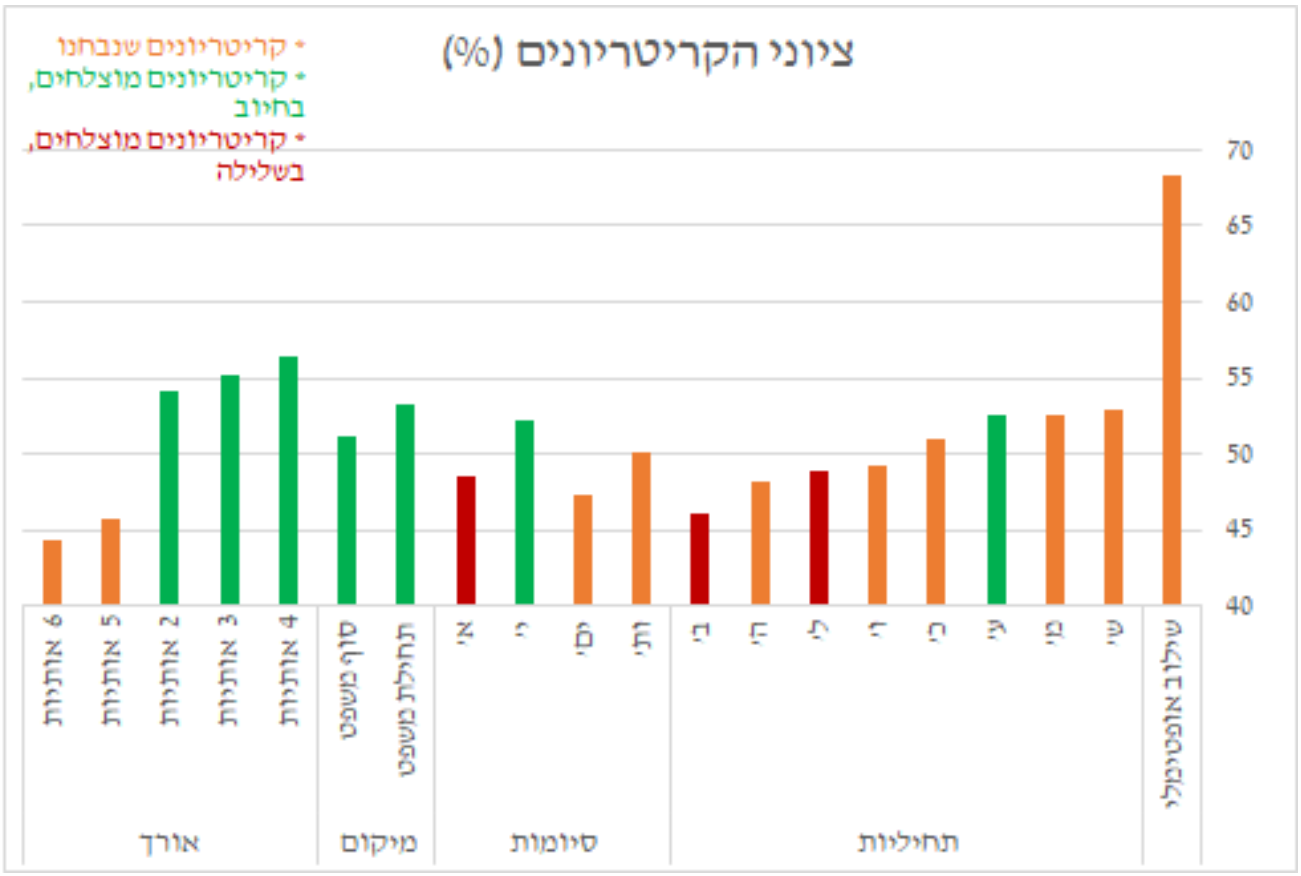
הקריטריונים (ביחד)	**הציון** (%)
תחילת משפט + סוף משפט + 2 אותיות + 3 אותיות + 4 אותיות - מתחיל בב' - מתחיל בל' + מתחיל בע' - נגמר באי + נגמר בי'	68.02

ולהלן ציוני הקריטריונים בגרף עמודות שמאפשר להשוות ביניהם באופן ויזואלי.

בכתום - הקריטריונים שנבחנו ולא נכללו בשילוב הקריטריונים האופטימלי.

בירוק - הקריטריונים שנכללו בשילוב האופטימלי כשמילוי שלהם מעלה את מספר הקריטריונים שנענו ב-1 (את המשתנה CritFilled).

באדום - הקריטריונים שנכללו בשילוב האופטימלי כשמילוי שלהם מוריד את מספר הקריטריונים שנענו ב-1.


ציוני הקריטריונים (%)
* קריטריונים שנבחנו
* קריטריונים מוצלחים, בחיוב
* קריטריונים מוצלחים, בשלילה
70
65
60
55
50
45
40
שילוב אופטימלי
שי'
מי'
עי'
כי'
רי'
לי'
הי'
בי'
ותי'
יסי'
י'
אי'
תחילת משפט
סוף משפט
4 אותיות
3 אותיות
2 אותיות
5 אותיות
6 אותיות
תחילות
סיומות
מיקום
אורך

מסקנות ודיון

יודגש ראשית שתוצאות הניסוי הן כיווניות - כלומר, אין להסיק מתוכן, לדוגמה, ש"מילה רב משמעית נוטה להתחיל באות ע"', אלא ש"מילים שמתחילות באות ע' נוטות להיות רב משמעיות". זאת מכיוון שרוב המילים שפותחות באות ע' בטקסטים עבריים הן המילה "על", שהיא רב משמעית, ומשמעות התוצאות היא למעשה שמילה שפותחת באות ע' נוטה להיות המילה "על".

אם כן, התשובה לשאלת החקר שהעלינו היא שעל פי הקורפוס המצומצם עליו נערך החקר, המילים הבאות נוטות להיות רב משמעיות: מילים עבריות שהן קצרות מאוד (2-4 אותיות), פותחות משפט או מסיימות אותו, לא פותחות באות בכל"ם (ובפרט, ב-ב' וב-ל'), מסתיימות בי' ולא מסתיימות באי'. כמו כן, באופן מובהק פחות, גם מילים שמסתיימות ב'ות', לא מסתיימות ב'ים', לא פותחות ב-ה' ומתחילות ב-מ', ו' ו-ש' גם נוטות להיות רב משמעיות. אבחנה זו היא מובהקת סטטיטית עם נתון t של 0.315 הנמוך מנתון ה-1.96 הנדרש בכדי להשיג מובהקות סטטיסטית (ראה נספח 1). עם זאת, שילוב הקריטריונים הוא כפי הנראה לא חד מספיק כדי לאפשר זיהוי מהימן של מילים רב משמעיות ובכך למלא את המטרה שלשמה נערך המחקר, דהיינו ייעול תהליך הפגת העמימות המורפולגית. חרף כך, האפשרות המוצעת במחקר היא מבוססת היטב תיאורטית, ומחקר עתידי בתחום שישפר את דיוק האפיון עשוי בהחלט לתרום לייעול התהליך המדובר.

כמו כן, ניתן גם לבחון את השערותינו בנוגע להימצאות הקריטריונים הבודדים:

תחיליות: כצפוי, אותיות בכל"ם הראו נטיות משתנות. האות מ' אפיינה במעט, באופן מפתיע, מילים רב משמעיות, האותיות ב' ו-ל' אפיינו דווקא מילים חד משמעיות, והאות כ' לא הראתה זיקה מובהקת לאף צד. ש' אכן אפיינה מילים רב משמעיות, כתואם את השערתנו שנטייתה של האות ש' תתאם את זו של הקריטריון "המילה בתחילת משפט" אך באופן מובהק מעט פחות. ו' החיבור לא הראתה נטייה מובהקת, ומכאן ש-ו' החיבור מופיעה יותר לפני האיבר האחרון בחלק כולל מאשר לפני מילת חיבור בין איברי משפט מאוחה.

סופיות: באופן שתואם את השערתנו, סיומות "ים" ו-"ות" הראו שתיהן על נוכחות מילה חד-משמעית באופן לא מאוד מובהק, כשסיומת "-ות" מובהקת יותר מסיומת "-ים". כמו כן, סיומת י' הראתה באופן די מובהק על נוכחות מילה רב משמעית וסיומת א' אכן הראתה על כך שהמילה היא חד משמעית.

מיקום במשפט: הן מילים בתחילת משפט והן מילים בסוף משפט נטו להיות רב משמעיות. בולטות הציון של קריטריוני המיקום מראה שממצא זה הוא מובהק יחסית, וכאמור, כל מובהקות בקריטריונים אלה נוגדת את השערתנו. ממצא זה נתמך ע"י העובדה שמילים בתחילת משפט נוטות להיות שמות עצם - מתוך 28 המשפטים בקורפוס, 12 פותחים בשם עצם (48.5% מן המשפטים).

אורך: מילים נפוצות רב משמעיות כגון 'את', 'עם' ו'אבל' הן קצרות מאוד ונפוצות מאוד. באופן שמאשש זאת, מילים קצרות (שאורכן קטן מ-5 אותיות) נוטות להיות רב משמעיות, ומילים ארוכות (שאורכן מ-5 אותיות ומעלה) נוטות להיות חד משמעיות.

שוב יצויין כי הקורפוס שעליו נערך המחקר הוא מצומצם יחסית (כ-500 מילים שנלקחו מתוך שלושה מקורות בלבד), ולכן את הנתון המספרי (68.02 אחוזי דיוק לגבי שילוב הקריטריונים האופטימלי) כמו גם את הקריטריונים תלויי המשפט, קריטריוני המיקום, יש לקחת בעירבון מוגבל. ניתן היה לשפר דיוק

זה באמצעות שימוש בקורפוס גדול יותר - בקטעי טקסט ארוכים יותר ובמגוון רב יותר שלהם. עם זאת, סוג אחד של הגדלה בא על חשבון השני: שימוש בקטעים ארוכים יותר מגביל את כמות הקטעים השונים בהם נעשה שימוש, ובכך מגביל את מגוונם ואת הייצוגיות של הקורפוס. שימוש במספר רב של קטעי טקסט מכריח אותנו להשתמש בכמות קטנה מכל "סוג" (משלב, תחום חיים, זהות כותב...) ובכך, שוב, להגביל את ייצוגיות הקורפוס.

כאמור, ניתן לייעל עוד את הניתוח על ידי מתן משקל שונה לכל קריטריון, מכיוון שכפי שניתן לראות בבירור, ישנם קריטריונים "מוצלחים" יותר ופחות. אם ייעשה דבר כזה בעתיד, מחקר נרחב מאוד על קורפוס רחב ומייצג יצטרך להיערך בכדי למצוא את הניקוד האופטימלי לכל קריטריון. מחקר זה עשוי להיערך בקלות על ידי רשתות נוירונים מלאכותיות (אלגוריתמים ללמידת מכונה שמדמים את מבנה המוח), שכן רשתות נוירונים אינן אלא אלגורתימים לאופטימיזציה של קריטריונים (גו ואחרים, 2009). רשת זו תוכל, כפי שהוזכר, אף לייצר קריטריונים מוצלחים יותר מאלה שהוצגו בעבודה זו, על חשבון הפשטות וה"חיצוניות" שלהם.

כמו כן, במהלך סיעור המוחין שהוביל לשאלת המחקר עלו מספר שאלות שנותרו ללא מענה. שאלות אלו תוכלנה לשמש כפתח למחקר עתידי בתחום האפיון השטחי של מילים רב משמעיות.

- מה היא היעילות הסיבוכית[6] של האלגוריתם הסטנדרטי להפגת עמימות מורפולוגית? של האלגוריתם החדש? של הפעלת הקריטריונים עצמה?
- האם האלגוריתם החדש אכן יעיל יותר מהישן? בכמה? מה הקשר בין מידת הייעול לציון של שילוב הקריטריונים האופטימלי?
- מה הקשר בין מידת הייעול של האלגוריתם החדש לכמות הקריטריונים? האם היה כדאי להסתפק בפחות קריטריונים?
- האם השילוב שנמצא הוא השילוב האופטימלי? האם השילוב האופטימלי יישתנה אם נשנה את הקורפוס? מהו השילוב האופטימלי האמיתי, כפי שייבחן על קורפוס מייצג וגדול מאוד? האם, בין הקריטריונים שלא נבחנו, קיים קריטריון יעיל יותר מאלה שנבחנו?
- האם קריטריונים שקיבלו ציונים טובים "בנפרד" יקבלו ציונים טובים "ביחד"? האם קיימים קריטריונים טובים שהשילוב ביניהם אינו טוב? האם קיימים קריטריונים לא טובים שהשילוב ביניהם טוב?
- מה הקשר בין כמות הקריטריונים שנענו על ידי כל מילה למספר הניתוחים המורפולוגיים שלה? האם האפיון של מילים עם מספר קריטריונים מוגדר (נגיד, ארבעה בדיוק) שונה מזה של מילים שהן פשוט רב משמעיות (כלומר, בעלות שני ניתוחים ומעלה)? מהו?

[6] נתון מתמטי של שגרות כשלכל שגרה סיבוכיות משלה. נתון זה מייצג את הזמן שייקח למחשב אופטימלי להפעיל את השגרה על קלט בגודל מסויים, כפונקציה של גודל הקלט. הסבר רחב יותר על הנושא ניתן למצוא בספרם של קורמן, ליזרסון, ריבט ושטיין (1990).

ביבליוגרפיה

1. Aronoff, M., & Fudeman, K. A. (2012). *What is morphology?* Chichester: Wiley-Blackwell.
2. Bar-Haim, R., Simaan, K., & Winter, Y. (2008). Part-of-speech tagging of Modern Hebrew text. *Natural Language Engineering*, *14*(2), 223–251. doi: 10.1017/s135132490700455x
3. Cormen, T. H., Leiserson, C. E., Rivest, R. L., & Stein, C. (n.d.). *Introduction to algorithms*.
4. Eliason, O. (2019, December 26). *Eich hichchadenu et hammamutot... Shuv* . Retrieved from https://davidson.weizmann.ac.il/online/sciencenews/איך-הכחדנו-את-הממותות-שוב
5. Gu, Xiao-Feng & Liu, Lin & Li, Jian-Ping & Huang, Yuan-Yuan & Lin, Jie. (2009). Data Classification based on Artificial Neural Networks. 223 - 226. 10.1109/ICACIA.2008.4770010.
6. Melnik, N., & Bottoinik , I. (2017). *Heker Ha'Safa: Yesodot Ve'yisumim*. Ra'anana, Israel.
7. Oxenberg, E. (2019, September 26). *Reakh Shel Yam*. Retrieved from https://davidson.weizmann.ac.il/online/askexpert/ריח-של-ים
8. Online, M. (2018, July 31). *Sirton ha-plastic ba-hasa*. Retrieved from https://www.maariv.co.il/news/israel/Article-653928
9. Schwarzwald, O. (2002). *Perakim be-morfologyah 'Ivrit*. Tel Aviv: ha-Universiṭah ha-petuḥah.

נספחים

נספח 1 - מבחן t על ציוני הקריטריונים

כדי לוודא שהציונים המושגים הם אכן מובהקים סטטיסטית, נריץ עליהם מבחן סטטיסטי המכונה "מבחן t". מבחנים מסוג זה בוחנים את המתאם בין שני משתנים אקראיים (במקרה שלנו, רשימת ניתוחי המחשב ורשימת הניתוחים של מורפיקס ומיל"ה) כאשר פיזור המשתנים סביב הממוצע שלהם (השונות) אינו ידוע.

מעשית, נייצר, דרך הקורפוס, שני טורים בתוכנת אקסל, האחד של הניתוחים שייצר המחשב בעזרת שילוב הקריטריונים האופטימלי, והשני של הניתוחים שיצרו מורפיקס ומיל"ה. לאחר מכן, נבדוק את המתאם בין שני הטורים: נייצר טור שלישי, בו העמודות הן 0 אם יש בשתי העמודות המתאימות בטור 1 ו-2 התאמה ו-0 אם אין. ההתאמה בין שני הטורים היא ממוצע הערכים הללו, כלומר, החלק של ההתאמות המוצלחות מתוך ההתאמות שנעשו. כמו כן, יחושב המתאם בין שני הטורים - זהו ערך, בין 1 ל-1- שמייצג (בערכו המוחלט) את מידת ההתאמה הלינארית בין שני טורי הנתונים (בניגוד להתאמה שנבדקת עבור צמדי נתונים), כאשר 1 מייצג התאמה מוחלטת, -1 מייצג התאמה הפוכה מוחלטת ו-0 מייצג היעדר התאמה. בטבלה להלן מודגם התהליך על המשפט הראשון בקורפוס (שאורכו 12 מילים):

מורפיקס	מחשב	האם ישנה התאמה?		ממוצע
1	1	TRUE	1	0.583
1	1	TRUE	1	מתאם
0	0	TRUE	1	0.169
1	1	TRUE	1	
0	0	TRUE	1	
1	0	FALSE	0	
1	0	FALSE	0	
0	0	TRUE	1	
0	1	FALSE	0	
0	0	TRUE	1	
0	1	FALSE	0	
0	1	FALSE	0	

לאחר מכן, נבצע באמצעות התוכנה את מבחן ה-t, שיחזיר שני נתונים שמעניינים אותנו: הראשון, t-stat, מייצג את היחס בין המידה בה הערכים חורגים זה מזה (השונות) לסטיית התקן שלהם. השני, t-Critical two-tail, מייצג את החלק של ההתפלגות של הערכים שנחשב למשמעותי, כלומר את החלק מה"זנב" של גרף הפעמון של הנותנים שנתעלם ממנו. אם הערך המוחלט של הנתון הראשון קטן משל השני, הושגה מובהקות סטטיסטית. בטבלה להלן מובאות תוצאות מבחן זה על המשפט הראשון. כפי שניתן לראות, ה-t-stat הוא אכן קטן (בערכו המוחלט) מה- t-Critical two-tail, ועל כן ההתאמה (0.583) והמתאם (0.169) הם מובהקים סטטיסטית.

	מורפיקס	מחשב
ממוצע:	0.416667	0.5
שונות:	0.265152	0.272727

תצפיות :	12	12
הפרש משוער :	0	
df	22	
t Stat	**-0.39361**	
P(T<=t) one-tail	0.348829	
t Critical one-tail	1.717144	
P(T<=t) two-tail	0.697657	
t Critical two-tail	**2.073873**	

וטבלת מבחן ה-t של כל הקורפוס :

	מורפיקס	מחשב	התאמה
ממוצע :	0.486	0.476	0.680
שונות :	0.250	0.250	מתאם
תצפיות :	504	504	0.360
הפרש משוער :	0		
df	1006		
t Stat	**0.315**		
P(T<=t) one-tail	0.376		
t Critical one-tail	1.65		
P(T<=t) two-tail	0.753		
t Critical two-tail	**1.96**		

גם בטבלה זו ערך ה t-stat הוא קטן (בערכו המוחלט) מה- t-Critical two-tail, ועל כן ההתאמה (0.680) והמתאם (0.360) הם מובהקים סטטיסטית - התוצאה שלנו, ששילוב הקריטריונים האופטימלי נותן דיוק רב יחסית של 68.02%, היא משמעותית.

נספח 2 - בוחן קריטריונים

בקטע קוד זה, שנכתב בשפת התכנות פייתון, נעשה שימוש בכדי לבחון יעילות קריטריונים. קוד זה לוקח משפט מנותח מראש, ומשווה אותו עם משפט שמנותח ע״י המחשב - מילה נחשבת ע״י המחשב לרב משמעית אם היא עונה על X קריטריונים, כמוחלט מראש בכל ניסוי. בעבודה זו הונח שקריטריון אחד מספיק; ניתן לשלב צורך למענה על מספר רב יותר של קריטריונים כשיהיה צורך במתן משקל לקריטריונים הללו, ניסוי שהושאר כפתח למחקר עתידי.

בקטע זה מוגדרות שלוש שגרות: WrdPerf, שמקבלת מילה בנוסף למידע על האם היא בתחילת משפט והאם היא בסוף משפט, ומחזירה 'אמת' אם המילה עומדת בקריטריון אחד לפחות ו'שקר' אם היא לא; SentPerf, שמקבלת משפט ומחזירה רשימה (״list״) בה כל איבר הוא אמת/שקר לפי ההפעלה של השגרה WrdPerf על המילה המתאימה במשפט; ו-Eval, שמקבלת משפט וגם רשימה של הניתוח הנכון שלו שנעשה באמצעות הכלי של מיל״ה ושל מורפיקס, ומחזירה את אחוז ההתאמה בין הניתוח הנכון לניתוח של המחשב - כמות הפעמים שהניתוח של המחשב תאם את הניתוח של מורפיקס חלקי כמות המילים במשפט, כפול מאה אחוז. לבסוף, קטע הקוד מחזיר את ממוצע הציונים (הפעלות של Eval על כל 28 המשפטים) עבור מערך קריטריונים מסויים.

כמו כן, בקוד מובא גם הקורפוס המחולק למשפטים (SentenceX) כמו גם הניתוחים הנכונים של מורפיקס ומיל״ה (RealSentPerfX)

```
def WrdPerf(word, isStart, isEnd):
    CritFilled = 0
    if isStart:
        CritFilled += 1
    if isEnd:
        CritFilled += 1
    if len(word) == 2:
        CritFilled += 1
    if len(word) == 3:
        CritFilled += 1
    if len(word) == 4:
        CritFilled += 1
    if word.endswith(״יי״):
        CritFilled += 1
    if word.endswith(״א״):
        CritFilled -= 1
    if word[0] == ״ב״:
        CritFilled -= 1
    if word[0] == ״ל״:
        CritFilled -= 1
    IsAmb = CritFilled >= 1
    return IsAmb
```

```
def SentPerf(sentence):
AreAmb = []
SplitSentence = sentence.split(" ")
for word in SplitSentence:
        IsStart = word == SplitSentence[0]
        IsEnd = word == SplitSentence[-1]
        AreAmb.append(WrdPerf(word,IsStart,IsEnd))
return AreAmb

def Eval(sentence,RealSentPerf):
SuccessfullEvals = 0
SentencePerformance = SentPerf(sentence)
for i in range(len(SentencePerformance)):
        if SentencePerformance[i] == RealSentPerf[i]:
        SuccessfullEvals += 1
return 100*SuccessfullEvals/len(SentencePerformance)

###This part uses MILA
##ריח של ים
Sentence1 = "בעיני רבים ההגדרה של החופשה המושלמת היא בטן גב ברצועת חוף אקזוטית"
RealSentencePerf1 =
[True,True,False,True,False,True,True,False,False,False,False,False]
Sentence2 = " המרכיבים ההכרחיים הם חול בין אצבעות הרגליים שמש מלטפת המיית הגלים ואוויר של ים"
RealSentencePerf2 =
[True,False,True,True,True,False,True,True,False,False,False,False,True,False]
Sentence3 = " אבל מהו בעצם אוויר של ים איך הוא מקבל את הניחוח הייחודי שלו והאם כדאי לנו לדעת מה אנחנו מכניסים לריאות כשאנחנו לוקחים נשימה עמוקה על החוף"
RealSentencePerf3 =
[True,False,True,False,True,False,False,True,True,True,False,True,True,True,True,Tr
ue,True,True,False,False,True,False,False,False,False,False,True,False]
Sentence4 = " כשמבקשים מאנשים לתאר את ריחו של הים התשובות הנפוצות כוללות בדרך כלל את המושגים רענן טרי מלוח או דגי בווריאציות שונות"
RealSentencePerf4 =
[True,False,False,True,False,True,False,False,True,False,True,True,True,True,True,Fa
lse,True,False,True,False,True,]
Sentence5 = " ריח הים נוצר מקוקטייל של כימיקלים רבים שמקורם בריקבון מוות ופירוק חיידקי עם קורטוב של מליחות אצות ואורגניזמים ימיים"
RealSentencePerf5 =
[False,False,True,False,True,False,True,True,False,False,False,True,True,False,True,
True,True,False,False]
```

Sentence6 = " לעיתים הריח יקושר לסביבה הקרובה יותר כמו הריח של קרם הגנה או של כלב רטוב לאחר טבילה בים"
RealSentencePerf6 =
[True,True,False,False,False,True,True,True,True,True,True,False,True,False,False,False,True,False,True]
Sentence7 = "יש מי שיפליגו על גלי הדימויים וידברו על ריח של חופש נעורים אנרגיה ובריאות"
RealSentencePerf7 =
[True,True,False,True,True,False,False,True,False,True,True,False,False,True]
Sentence8 = " חובבי המילה הכתובה בטח ייזכרו במשורר או בסופר האהובים עליהם יוצקים בריח הים רומנטיקה ואהבה"
RealSentencePerf8 =
[False,False,True,True,False,True,False,True,True,True,True,False,False,False,True]
Sentence9 = "פחות סביר שתשמעו את התשובה המדעית המלאה ויש לזה סיבה טובה"
RealSentencePerf9 = [True,True,False,True,False,False,True,True,False,False,True]
Sentence10 = " ריח הים נוצר מקוקטייל של כימיקלים רבים שמקורם בריקבון מוות ופירוק חיידקי עם קורטוב של מליחות אצות ואורגניזמים ימיים"
RealSentencePerf10 =
[False,False,True,False,True,False,True,True,False,False,False,False,True,False,True,True,True,False,False]
Sentence11 = "לא נשמע רומנטי ורענן כמו רוב השירים שנכתבו על הים אבל זה מה שיש"
RealSentencePerf11 =
[False,True,False,True,True,True,False,False,True,False,True,False,True,False]
Sentence12 = " אז כדי להבין קצת יותר מאיפה מגיע הריח ומה מרכיב אותו נבחן כמה מהמולקולות היותר נפוצות שמתעופפות במשבי הבריזה"
RealSentencePerf12 =
[True,True,False,True,True,False,False,True,False,True,False,True,True,False,False,True,False,False,True]
###This part uses Morfix
Sentence13 = " מדובר במולקולה קטנה אך מסריחה שריחה מתואר כדומה לזה של כרוב או אספרגוס מבושל"
RealSentencePerf13 =
[True,False,True,False,True,False,True,True,False,True,True,False,False,True]
Sentence14 = "אבל לא מדובר בחומר שריחו בהכרח דוחה אותנו"
RealSentencePerf14 = [True,False,True,False,False,False,True,False]
Sentence15 = " בריכוזים מסוימים תוכלו למצוא אותו ביין או בירה במגוון ירקות ופירות לרבות עגבניות ומנגו והוא תורם גם לניחוח של גבינות קשות מיושנות ופטריות כמהין"
RealSentencePerf15 =
[False,True,False,False,True,False,False,False,True,False,False,True,True,False,False,True,False,False,True,False,True,False,False,False,]
Sentence16 = " טכנולוגי מזון אף משתמשים בו כדי להעניק טעמי בשר דגים ביצים חמאה פירות וירקות למוצרי מזון מעובדים"

RealSentencePerf16 =
[True,False,True,True,False,True,False,True,True,True,False,False,False,False,False,False,True]
סרטון הפלסטיק בחסה##
Sentence17 = " בימים האחרונים סוערת הרשת הישראלית בעקבות סרטון בו מראים שבחסה שאנו אוכלים קיימת שכבת פלסטיק שניתן ממש לקלף אותה אחרי שמטביעים את החסה במים רותחים"
RealSentencePerf17 =
[False,False,True,True,True,True,False,False,False,False,True,True,True,True,False,False,True,False,True,False,False,True,False,False,True]
Sentence18 = " אולם בפוסט ויראלי שעלה בעמוד הפייסבוק סטטוסים מצייצים נראה כי כל עניין הפלסטיק בחסה יצא מפרופורציות ומדובר בסך הכל באפידרמיס שכבה שמגנה על העלה ונמצאת גם בעורם של בני האדם"
RealSentencePerf18 =
[True,False,False,True,False,False,False,False,True,False,False,True,False,True,False,False,True,False,False,True,True,True,True,False,False,False,True,False,False]
Sentence19 = " מצב החקלאות גם ככה בקריסה נכתתב בפוסט הסותר את טענת הפלסטיק בחסה"
RealSentencePerf19 =
[True,False,False,False,False,False,False,True,True,True,False,False]
Sentence20 = " מספיקה לנו הממשלה שדואגת לחסל לאט לאט אבל בטוח את ענף החקלאות הישראלי"
RealSentencePerf20 =
[True,True,False,True,False,True,True,True,True,True,True,False,True]
Sentence21 = "אין לנו צורך באנשים ללא כל ידע שינסו לחסל את החקלאות"
RealSentencePerf21 = [True,True,True,False,False,False,True,True,False,True,False]
Sentence22 = " אלו אנשים שידם קלה על המקלדת אנשים שלא חושבים פעמיים מי נמצא בצד השני"
RealSentencePerf22 =
[True,True,True,True,True,False,False,False,False,False,True,False,False,True]
Sentence23 = " כותבת הפוסט הישראלית שמספרת כי אביה הוא בעל משק חקלאי כבר שלושים שנים ביקשה בפוסט שאם חשובה לכם מדינת ישראל והחקלאות הישראלית התעלמו מהסרטון תסביאו לחברים שלכם מה זה אומר ותזכירו לכולם שבישראל רק חקלאות ישראלית"
RealSentencePerf23 =
[True,False,True,False,False,False,False,False,True,True,True,False,False,False,True,False,False,True,False,False,False,True,False,False,False,False,False,False,False,False,True,True,False,False,True,False,True]
Sentence24 = " ותמשיכו לאכול חסה לא מסרטנת מלאה בברזל טובה לעיכול ועוד מלא סגולות ששמורות רק לה ולא לאף ירק אחר"

RealSentencePerf24 =
[False,True,True,False,False,True,False,True,False,False,True,True,True,True,False,False,True,True,True]
Sentence25 = "לפוסט צורף סרטון של אביה בו הוא מסביר שנעשה לחסה עוול גדול"
RealSentencePerf25 =
[False,True,False,True,False,False,False,True,False,True,False,True]
איך הכחדנו את הממותות... שוב##
Sentence26 = "בערבות סיביר פורחת בשנים האחרונות תעשייה יוצאת דופן באופייה"
RealSentencePerf26 = [True,False,True,False,True,False,True,False,False]
Sentence27 = " חטים קדומים של ממותות ששרידים רבים מהן השתמו בקרקעות הקפואות נחפרים בהמוניהם מתוך האדמה כדי לענות על הביקוש הרב לשנהב בסין שם תכשיטים וחפצי אומנותג מגולפים משנהב הפכו סמל סטטוס לעשירים"
RealSentencePerf27 =
[False,True,True,False,False,True,True,False,False,False,False,False,True,False,True,True,True,False,True,False,False,True,False,True,True,True,False,False,True,False,True]
Sentence28 = " הסחר שהיקפו עומד על יותר מחמש מאות טונות בשנה מתנהל ברובו הרחק מעיני הרשויות והחוק ועוזר לפרנס את תושביו של אחד האזורים העניים ביותר ברוסיה"
RealSentencePerf28 =
[False,False,True,False,True,True,False,False,True,False,False,True,True,False,False,True,True,True,False,True,True,True,True,False,True]

Eval1 = Eval(Sentence1, RealSentencePerf1)
Eval2 = Eval(Sentence2, RealSentencePerf2)
Eval3 = Eval(Sentence3, RealSentencePerf3)
Eval4 = Eval(Sentence4, RealSentencePerf4)
Eval5 = Eval(Sentence5, RealSentencePerf5)
Eval6 = Eval(Sentence6, RealSentencePerf6)
Eval7 = Eval(Sentence7, RealSentencePerf7)
Eval8 = Eval(Sentence8, RealSentencePerf8)
Eval9 = Eval(Sentence9, RealSentencePerf9)
Eval10 = Eval(Sentence10, RealSentencePerf10)
Eval11 = Eval(Sentence11, RealSentencePerf11)
Eval12 = Eval(Sentence12, RealSentencePerf12)
Eval13 = Eval(Sentence13, RealSentencePerf13)
Eval14 = Eval(Sentence14, RealSentencePerf14)
Eval15 = Eval(Sentence15, RealSentencePerf15)
Eval16 = Eval(Sentence16, RealSentencePerf16)
Eval17 = Eval(Sentence17, RealSentencePerf17)
Eval18 = Eval(Sentence18, RealSentencePerf18)

```
Eval19 = Eval(Sentence19, RealSentencePerf19)
Eval20 = Eval(Sentence20, RealSentencePerf20)
Eval21 = Eval(Sentence21, RealSentencePerf21)
Eval22 = Eval(Sentence22, RealSentencePerf22)
Eval23 = Eval(Sentence23, RealSentencePerf23)
Eval24 = Eval(Sentence24, RealSentencePerf24)
Eval25 = Eval(Sentence25, RealSentencePerf25)
Eval26 = Eval(Sentence26, RealSentencePerf26)
Eval27 = Eval(Sentence27, RealSentencePerf27)
Eval28 = Eval(Sentence28, RealSentencePerf28)
Evals =
[Eval1,Eval2,Eval3,Eval4,Eval5,Eval6,Eval7,Eval8,Eval9,Eval10,Eval11,Eval12,Eval13
,Eval14,Eval15,Eval16,Eval17,Eval18,Eval19,Eval20,Eval21,Eval22,Eval23,Eval24,Eva
l25,Eval26,Eval27,Eval28]

def avg(numbers):
return sum(numbers)/len(numbers)

print (avg(Evals))
```