זיהוי אוטומטי של מסגרות ההצרכה של הפועל בעברית

חנה פדידה

זיהוי אוטומטי של מסגרות ההצרכה של הפועל בעברית

חיבור על מחקר

לשם מילוי חלקי של הדרישות לקבלת תואר מגיסטר למדעים במדעי המחשב

חנה פדידה

הוגש לסנט הטכניון – מכון טכנולוגי לישראל

סיוון התשע"ב חיפה יוני 2012

המחקר נעשה בהנחייתם של פרופ׳ אלון איתי ופרופ׳ שולי וינטנר בפקולטה למדעי המחשב

אני מודה לטכניון על התמיכה הכספית הנדיבה בהשתלמותי

תוכן העניינים

רשימת סמלים וקיצורים

BHT	Binomial Hypothesis Test
COM	Complement
DEP	Dependency
FN	False Negative
FP	False Positive
LLR	Log Likelihood Ratio
MLE	Maximum Likelihood Estimation
MWV	Multi-Word Verbs
OBJ	Object
PDT	Prague Dependency TreeBank
PMI	Pointwise Mutual Information
PP	Prepositional Phrase
RF	Raw Frequency
TN	True Negative
TP	True Positive
מיל״ה	מרכז ידע לתקשוב בשפה העברית

רשימת טבלאות

רשימת איורים

תקציר

הפועל הוא העומד בבסיס המבע, ובו מתרכז תוכנו העיקרי של המשפט. לא תמיד יש לפועל קיום שלם כיחידה בודדת, וישנם פעלים רבים הדורשים שמידע נוסף יצטרף אליהם, על מנת שמשמעותם תהיה שלמה ומלאה. לדוגמה, הפועל **הגן** אינו שלם ללא צירוף יחס עם מילת היחס **על**, הפועל **הבחין** מצריך צירוף יחס עם מילת היחס **ב**. המידע הנוסף הזה הדרוש לפועל הוא *מסגרת ההצרכה*, מונח אותו טבע חומסקי (1965), ובתופעה זו מתרכז מחקרנו.

מחקרים רבים נעשו על מציאת מסגרת ההצרכה באופן אוטומטי עבור שפות שונות כמו אנגלית (Korhonen et al., 2006) וצרפתית (Messiant et al., 2008), ואף עבור שפות שמיות כערבית (Attia et al., 2011), אך העברית נותרה מאחור. למיטב ידיעתנו, זוהי העבודה הראשונה העוסקת בזיהוי אוטומטי של מסגרת ההצרכה של הפועל בשפה העברית.

ידיעת מסגרת ההצרכה של הפועל היא בעלת חשיבות רבה למגוון יישומים, שהחשוב ביניהם הוא ניתוח תחבירי. מחקרים הראו כי כ-50% מהניתוחים השגויים שמתקבלים בניתוח משפטים אוטומטי הם כתוצאה מחוסר מידע על מסגרות ההצרכה של הפעלים (Carroll, 1998). ואכן, הוספת המידע הזה למנתח תחבירי הביאה לשיפור של ממש בניתוחים המופקים על ידו (Zeman, 2002; Carroll, Minnen & Briscoe, 1998).

מחקרנו הינו מחקר מבוסס ומונע קורפוס. אנו משתמשים בתכונות הייחודיות של מסגרת ההצרכה המופיעות במשלב הנחקר (הקורפורה בהם אנו משתמשים הם ברובם טקסטים עיתונאיים) על מנת לפתח טכניקה המתרכזת באזור הקרוב ביותר לפועל וממנו דולה ולומדת מידע רב על הפועל (פרק 4).

במסגרת החיבור נסקור ראשית את תופעת מסגרת ההצרכה וביטוייה בשפה העברית וכן את ייחודיות העברית לעומת שפות אחרות (פרק 1), נסקור את העבודות החשובות בתחום ואת הגישות השונות ללמידת מסגרות ההצרכה (פרק 2) ונציג את מתודולוגיית המחקר (פרק 3). החלק השני של העבודה יתרכז בתהליך הלמידה. בחלק זה נציג את שתי השיטות השונות אותן בחנו ללמידת מסגרת ההצרכה של הפועל - למידת מסגרת ההצרכה מקורפוס מתויג מורפולוגית (פרק 5) ולמידת מסגרת ההצרכה מקורפוס המנותח ניתוח תחבירי (פרק 7).

מטרת מחקרנו היא להעמיד לקסיקון פועלי, הקיים בשפות אחרות, כדוגמת VALEX לאנגלית (Korhonen et al., 2006) או LexSchem לצרפתית (Messiant et al., 2008), אשר מכיל בנוסף למסגרות ההצרכה השונות לכל פועל, גם מידע סטטיסטי על היקרויות הפועל. הלקסיקון הפועלי אותו יצרנו מכיל מידע עבור 3,393 פעלים שונים, וכולל 20,829 זוגות פועל-מסגרת הצרכה (סעיף 5.3). במסגרת פרקים אלו נציג גם שלוש שיטות ששימשו להערכת איכות תוצאותינו: 1. הערכה ישירה באמצעות קבוצת בדיקה (סעיף 6.1) 2. שימוש בתוצאות כדי לפתור את בעיית הצמדת צירופי יחס (סעיף 6.3), 3. שימוש בתוצאות לשיפור האיכות של תרגום אוטומטי (סעיף 6.4). התוצאות שהתקבלו מראות שיפור ניכר בהצמדת צירוף היחס (הפחתה של 28.85% בשיעור הטעויות) ובתרגום אוטומטי מערבית לעברית (שיפור במדד BLUE מ-32.5 ל-37, ובמדד METEOR מ-52.6 ל-56).

אנו מקווים כי תוצאות מחקרנו ישמשו את הקהילייה המדעית העוסקת בשפה העברית, ויהוו בסיס למחקר פורה ונרחב של הפועל.

פרק 1

למידת מסגרות ההצרכה של הפועל

בפרק זה נציג את הרקע התיאורטי ואת המוטיבציה לעבודתנו. במסגרת פרק זה נסקור את התופעות הלשוניות של מסגרת ההצרכה וביטוייה בשפה העברית (סעיף 1.1), נציג את חשיבותה הרבה ושימושיה השונים (סעיף 1.2), ולבסוף נציג את מטרת מחקרנו (סעיף 1.3).

1.1 מסגרת ההצרכה של הפועל

הפועל הוא האלמנט המרכזי במשפט. הוא מבטא את תוכנו העיקרי של המשפט, ומורחב על ידי ארגומנטים שונים המוסיפים מידע על הפעולה. ארגומנטים כאלו יכולים למלא תפקידים של מבצע הפעולה, מושא הפעולה, אופן הפעולה, וכן להוסיף מידע לגבי מקום הפעולה וזמנה. אך לפועל אין תמיד קיום שלם כיחידה בודדת, וישנם פעלים רבים הדורשים שמידע נוסף יצטרף אליהם, על מנת שמשמעותם תהיה שלמה ומלאה. המידע הנוסף הזה הוא *מסגרת ההצרכה* (*subcategorization frame*), מונח אותו טבע חומסקי (1965). הדיון הבלשני על מסגרות ההצרכה מבחין בין מסגרת ההצרכה **התחבירית** למסגרת ההצרכה **הסמנטית** (בורוכובסקי-בר אבא, 2001). מסגרת ההצרכה התחבירית היא זו המגדירה את סוגי הארגומנטים שהפרדיקט הפועלי מקבל מבחינה תחבירית, ואילו מסגרת ההצרכה הסמנטית קשורה למשמעות הפועל, ובה מוגדר תפקידו התמטי של כל ארגומנט והיא כוללת גם את ההתניות הסמנטיות על הארגומנטים. לדוגמה, הפועל **אכל** דורש כי מבצע הפעולה יהיה שם עצם חי, המסוגל לבצע את פעולת האכילה, וכן כי הדבר עליו מבוצעת הפעולה יהיה אכיל.

מסגרת ההצרכה הסמנטית חשובה לכל תהליך של עיבוד שפה טבעית, אך קשה לזיהוי באמצעים אוטומטיים, ובמיוחד עבור עברית, בה לא קיימים כמעט כלים המאפשרים לבצע עיבוד סמנטי. עקב כך, במחקר זה אנו בוחנים רק את מסגרות ההצרכה התחביריות של הפועל, ומכאן ואילך הביטוי **מסגרת הצרכה** יתייחס למסגרת התחבירית.

הארגומנטים היכולים להופיע כמשלים לפועל רבים, ואנו בחרנו להתמקד במחקר בארבעה סוגי משלימים עיקריים. ההבחנה בין המשלימים היא על פי חלקי הדיבר שלהם ומבנם הצורני בלבד, ולא על פי תפקידם הפונקציונלי.

ארבעת סוגי המשלימים הם:

1. צירוף שמני

(א) החייל **קיבל טיפול רפואי** במקום. (גל"צ, 11.2.11)

(ב) שופטת בית הדין האזורי לעבודה **זימנה את יו"ר ההסתדרות** להמשך הדיון. (הארץ, 20.2.12)

בעברית קיימים שני ביטויים שונים להשלמה של צירוף שמני: הצטרפות עם המילית 'את', והצטרפות בלעדיה. הוספת המילית 'את' תלויה בצירוף השמני: צירוף שמני שהינו מיודע, שם פרטי או כינוי חבור דורש הופעת 'את' לפניו, ובכל יתר המקרים המילית 'את' לא יכולה להופיע.

2. צירוף יחס

(א) ענן האפר הוולקני **מנע מ**-440 טיסות להמריא מאירלנד. (גלובס, 4.5.11)

(ב) עובדי משרד הפנים **פתחו ב**עיצומים. (הארץ, 20.2.11)

צירוף היחס בנוי ממילת יחס + צירוף שמני, ומילת היחס היא הקשורה קשר של הצרכה לפועל.

3. פסוקית

(א) טהראן **הודיעה כי** תפסיק למכור נפט לשש מדינות אירופיות. (הארץ, 19.2.12)

הפסוקיות המתקשרות לפועל ומהוות משלים מאופיינות בדרך כלל במיליות הקישור '**ש**' ו'**כי**', אולם קיימים גם מופעים אחרים כדוגמת דיבור ישיר ופסוקית הפותחת במילת שאלה.

4. שם פועל

(א) אלפי אנשים **נאלצו לעזוב** את המתחם. (ערוץ 7, 12.7.05)

השלמות על ידי צירוף שמני, פסוקית ושם פועל לעיתים מתחלפות זו בזו כבפועל **רצה**, המתיר את שלוש ההשלמות:

(א) מאבטח כלא פלסטינית כי **רצה לצאת** איתה. (הארץ, 7.2.12)

(ב) אלישע **רצה מזכרת** מקפריסין. (הארץ, 2.12.11)

(ג) לא **רצינו** שאנשים ידאגו שהמקום עלול להיגמר להם. (מעריב, 1.4.2005)

אך קיימים מקרים בהם לא כל המימושים ייתכנו:

(א) שרה **מגבשת תביעה** נגד עוזרת הבית. (הארץ, 20.2.12) – הפועל **מגבש** מקבל צירוף שמני אך לא פסוקית או שם פועל

(ב) העיתון **דיווח** ב-1986 **ש**גריימס מת בלוס אנג'לס שנתיים קודם לכן. (הארץ, 19.2.12) – הפועל **דיווח** מקבל פסוקית אך לא שם פועל

(ג) **החלטתי להקדיש** חודש מחיי לחיפוש אחר קונטרבס בשבילו. (הארץ, 19.2.12) – הפועל **החליט** מקבל שם פועל או פסוקית אך לא צירוף שמני

עיון בבנק העצים (3.2) מעלה כי מתוך 704 פעלים המקבלים צירוף שמני, פסוקית או שם פועל, רק כ-2% מקבלים את כל שלוש ההשלמות, וכ-13% מקבלים שתי השלמות אפשריות.

כמו כן, גם כאשר פועל יכול לקבל את שלוש ההשלמות, ייתכן שתהיה לו העדפה ברורה רק לאחת מהן, כמו הפועל **רצה**, שאמנם כפי שראינו יכול לקבל מספר השלמות, אך ברוב מופעיו הוא מופיע עם שם פועל. בשל כך אנו מבחינים בין הצירוף השמני, הפסוקית ושם הפועל, ולא מתייחסים אליהם כיחידת השלמה אחת, אלא כשלוש יחידות השלמה שונות.

מסגרת ההצרכה לא חייבת להכיל ארגומנט אחד בלבד, וישנן מסגרות הצרכה רחבות יותר, למשל, במשפטים הבאים:

(א) לא **התחשק לי לחפש** את מקומי בעולמות אחרים. (הארץ, 19.2.12)

(ב) התמזה **העניקה לי נקודת מבט** ייחודית. (הארץ, 19.2.12)

מסגרות ההצרכה של הפעלים **התחשק** ו-**העניק** מורכבות משני ארגומנטים. במשפט (א) המסגרת היא **התחשק + ל + שם פועל**, ובמשפט (ב) היא **העניק + ל + צירוף שמני**.

מספר הארגומנטים שמכילה מסגרת ההצרכה הוגדר על ידי רובינשטיין כ**ערכיות תחבירית** (רובינשטיין, תשל"ט). בין הארגומנטים איננו כוללים את הנושא, מכיוון שהנושא הוא חלק ממסגרת ההצרכה של כל פועל. ערכיות הפועל יכולה לנוע בין אפס ארגומנטים משלימים:

(א) **מת** - **מת** המלחין ומבקר המוזיקה בנימין ברעם. (הארץ, 19.2.12)

(ב) **התרסק** - מטוס ועליו 47 בני אדם **התרסק**. (הארץ, 13.9.10)

ועד לשלושה ארגומנטים משלימים:

(ג) **התערב** - בחודש שעבר **התערב** מאסק **על** מיליון דולר **עם** כתב הרכב של "וול סטריט ג'ורנל" **כי** הרכב ייצא במועדו המתוכנן. (דה מרקר, 9.10.11)

למידת מסגרת ההצרכה של הפועל היא משימה קשה. ראשית, מלבד מסגרת ההצרכה סובבים את הפועל משלימים שונים שאינם מוצרכים, והם בבחינת תוספות שאינן הכרחיות. דוגמאות לכך הם ביטויי המקום והזמן:

(א) בשנתיים האחרונות **אכלתי ארוחת ערב** במסעדות חמישה ערבים בשבוע. (הארץ, 16.10.11)

(ב) תופעה שכיחה אחרת היא קושי להיזכר מה **אכלתי** בבוקר. (הארץ, 1.3.05)

צירוף היחס "במסעדות" במשפט (א) אינו משלים מוצרך לפועל, אלא הוא תיאור מקום המלווה את פעולת האכילה, אך קיומה עומד גם ללא התיאור. גם הצירוף "בבוקר" שבמשפט (ב) העוסק בזמן הפעולה אינו הכרחי והוא תיאורי. לעומת זאת, קיימים פעלים עבורם ביטויי המקום או הזמן הם מוצרכים כדוגמת הפועל **גר** או **התפרץ**:

(ג) **גרתי** בדירת חדר וחצי שכורה **ב**רמת גן. (הארץ, 12.2.12)

(ד) בסוף הראיון ילד קטן **התפרץ לח**דר. (הארץ, 20.2.12)

מלבד ביטויי המקום והזמן, גם תיאורי האופן קשורים לפועל, אך אינם מוצרכים לו, כבמשפט:

(ה) הילד מלהג בקדחתנות על כך שהוא מוכרח למצוא מישהו. (הארץ, 20.2.12)

בו תיאור האופן "בקדחתנות" קשור לפועל **מלהג**, אך מילת היחס 'ב' כלל אינה מהווה חלק ממסגרת ההצרכה של הפועל.

למרות שהתופעה שכיחה, ההבחנה בין משלים מוצרך למשלים שאינו מוצרך היא קשה. אזר (תשל"ב) סוקר את ההבחנות שהיו רווחות בין הבלשנים בשנות ה-60 לזיהוי משלים מוצרך, ומראה כי אין תמימות דעים בקריטריון ההבחנה, ולכן מחזק את הטענה בדבר הקושי להבחין ביניהם. מסקנתו של אזר היא כי:

... **נוסף על כך ראינו שלא תמיד אפשר להבחין ביניהם, וחילוקי הדעות מרובים**. מוטב איפוא, שלא לקבוע מסמרות בנידון ובייחוד לא בחיבור מילוני...

מכיוון שהמטרה במחקר זה הינה לפתח שיטות אוטומטיות לזיהוי מסגרת ההצרכה, ואין קריטריונים ברורים להבחנה בין המשלימים המוצרכים לתיאורים, בחרנו להתעלם מההבחנה בין מושאים לתיאורים, ולכלול את כולם תחת אותה מטרייה של הצרכה, כשהדגש הוא סטטיסטי – היקרות תכופה של פועל ומשלים מרמזת לקשר שקיים ביניהם, ואיננו קובעים את אופי הקשר הזה.

מלבד ההכרעה בדבר ההצרכה, מציאת מסגרת ההצרכה השלמה והכוללת אף היא משימה לא פשוטה. קיימים מופעים רבים של פועל עם מסגרת הצרכה חלקית בלבד או דווקא רחבה יותר, בשל סיבות שונות. ללמידה אוטומטית, המתבססת על דוגמאות בלבד, הדבר יכול להפריע. סיבה אחת למימוש צר יותר הוא כי המשלים מובן מתוך הפעולה, לדוגמה:

(א) אלמנת הכבאי **ילדה** אתמול **בת**. (הארץ, 29.12.10)

למרות שהפועל **ילד** דורש משלים שמני, המשפט היה תקין לחלוטין גם אם המשלים "בת" לא היה מופיע:

(ב) אלמנת הכבאי **ילדה** אתמול.

מסגרת רחבה יותר יכולה להופיע למשל במקרה של מושא פנימי - הפועל עצמו אינו דורש משלים שמני, אך על מנת להוסיף תיאור מסוים לפועל המופע מורחב על ידי המושא הפנימי, שהוא בדרך כלל שם פעולה הנגזר משורש הפועל. דוגמה לכך ניתן לראות בפועל **ישן**. מסגרת ההצרכה של הפועל **ישן** היא ריקה, שכן הוא אינו דורש משלים, אך ישנם מופעים של הפועל בהם הפעולה מורחבת למשל במשפט:

(ג) גלעד בכלאו ועם ישראל **ישן את שנתו** בלילות. (הארץ, 18.9.10)

בנוסף, לפועל אין בהכרח מסגרת הצרכה אחת בלבד, אלא לעיתים ישנן מספר מסגרות הצרכה שונות, שמלבד השונות התחבירית יכולות לגלם בתוכן שונות סמנטית. ניתן לראות זאת בפעלים הבאים:

1 **קינא** - (א) הוא היה **מקנא לה** כאילו היא רכושו הפרטי.
(ב) לפני כמה שנים **קינאנו** יותר **ב**עובדי ההיי-טק.

2 **ניצח** - (א) הטניסאי הצ'יליאני **ניצח את** הישראלי.
(ב) נשיא קוסטה ריקה **ניצח על** תהליך השלום.

בלמידת מסגרות ההצרכה יש לדעת להבחין בין מסגרות ההצרכה השונות השייכות לאותו פועל, כאשר מראש החיפוש הוא אחר יותר ממסגרת הצרכה אחת. הדבר קשה בעיקר לאור העובדה כי לא ניתן לדעת האם מופע חדש של פועל עם משלים הוא מופע של מסגרת הצרכה חדשה או שמא מופע של תיאור, או אולי מימוש לא שלם של מסגרת ההצרכה.

הסדר בין המשלימים בתוך מסגרת ההצרכה גם הוא מתעתע בהכרעה, שכן הוא חופשי יחסית (Belletti & Shlonsky, 1995), ולכן ייתכנו מופעים של הפועל ומסגרתו בסדרים שונים. למרות זאת, קיימים מקרים בהם הסדר הוא קבוע, כדוגמת ביטויים קפואים כמו:

1 א) העמיד אותו על טעותו

ב) *העמיד על טעותו אותו

או סדר שהוא תוצאה של תלות בשם העצם המהווה את גרעינו של המשלים. בעברית נפוצה התופעה של התקרבות צירוף היחס בנטיית כינוי הגוף לפועל, לעומת משלימים אחרים (Berman Aronson, 1978), כדוגמת:

2 א) **העמיד את** הילד **על** הכיסא

ב) **העמיד על** הכיסא **את** הילד

אך:

3 א) **העמיד אותו על** הכיסא

ב) * **העמיד על** הכיסא **אותו**

תופעות אלו אינן ייחודית רק לעברית, והן נפוצות גם בשפות אחרות (Ross, 1967), אך הן עדיין מהוות מכשול בבואנו למצוא את מסגרת ההצרכה.

בשל כל התופעות לעיל ניתן לראות כי המשימה אותה הצבנו לעצמנו אינה קלה, ומהווה אתגר.

1.2 חשיבות מציאת מסגרת ההצרכה ושימושיה

לידיעת מסגרת ההצרכה של הפועל חשיבות רבה בעיבוד שפות טבעיות. על פי Carroll (1998), 50% מהניתוחים השגויים שמתקבלים בניתוח משפטים אוטומטי הם כתוצאה מחוסר מידע על מסגרות ההצרכה של הפעלים. הסיבה לכך היא שקיימת עמימות רבה לגבי הקשרים הנכונים לפועל, והוספת המידע על מסגרת ההצרכה יכולה לצמצם אותה. ואכן, עבודות שנעשו על שילוב מסגרות ההצרכה של הפועל במנתח הראו שיפור בתוצאות הניתוח. Zeman (2002), לדוגמה, השתמש בהוספת המידע על מסגרת ההצרכה של הפועל שנמצא בלקסיקון שפיתחו Sarkar ו-Zeman (2000) לשיפור מנתח תלויות סטטיסטי. Zeman השתמש הן במסגרת כולה והן בכל אחד מרכיביה בנפרד על מנת לקבל את ההחלטה לגבי התלויות, והראה שיפור בדיוק (accuracy) מ-79.9 ל-82.1.

עבודה נוספת שהראתה שימוש במסגרת ההצרכה בניתוח תחבירי, היא עבודתם של Carroll, Minnen ו-Briscoe (1998) אשר בחנו את השפעת הוספת המידע על מסגרות ההצרכה, כולל שכיחויותיהן של המסגרות עם הפועל, למנתח תחבירי סטטיסטי. השימוש במסגרת ההצרכה של הפועל נעשה בסופו של תהליך הניתוח, על מנת להכריע בין עצי ניתוח שונים. עבודתם משפרת את הדיוק (precision) מ-79.2 ל-88.8.

שימוש נוסף למסגרת ההצרכה הוא סיווג פעלים (Schule im Walde & Brew, 2002). Schule im Walde ו-Brew השתמשו במידע התחבירי של מסגרות ההצרכה על מנת לסווג פעלים בגרמנית למחלקות סמנטיות. כמו כן ניתן להשתמש במידע לתמצות של טקסטים (Surdeanu et al., 2003), ולשיפור תרגום אוטומטי (Hajič et al., 2004).

קיומו של לקסיקון פועלי המכיל מידע על מסגרות ההצרכה יכול להיות לעזר גם בתחומים שונים של הבלשנות כמו פסיכולינגויסטיקה. הידע המוקדם על מסגרת ההצרכה משפיע על תהליך עיבוד המשפט וניתוחו, ומשפיע על מהירות הקריאה, הפקת משפטים, ועוד (Garnsey et al., 1997). שימוש בנתונים על מסגרת ההצרכה יכול לשמש לעריכת ניסויים בתחום, או ללמידת תופעות לשוניות שונות (Lapata et al., 2001 ;Baldewein & Keller, 2004). Lapata, Keller ו-Schulte im Walde (2001), לדוגמה, משתמשים במסגרות הצרכה הנלמדות מתוך קורפוס על מנת לבדוק ולכמת אילו תכונות משפיעות בתהליך עיבוד המשפט שנעשה במוח האנושי.

1.3 מטרת העבודה

מתוך הכרה בחשיבות המידע על מסגרות ההצרכה לתחום הבלשנות החישובית, ובשל העובדה כי השפה העברית דלה עדיין בכלים חישוביים, מטרת מחקרנו היא להעשיר את השפה העברית בעוד כלי חיוני וחשוב. אנו בונים לקסיקון פועלי מקיף המכיל מידע סטטיסטי על מסגרות ההצרכה השונות לפועל, כך שיוכל להוות כלי עזר לחוקרים הן בתחום הבלשנות החישובית, והן בתחומים אחרים של הבלשנות. לקסיקון הפועל שלנו יהיה נגיש בזמן הקרוב לקהילייה המדעית.

פרק 2

סקר ספרות

זיהוי מסגרות ההצרכה של הפועל ולמידתן באופן אוטומטי היא משימה שהעיסוק בה החל בתחילת שנות ה-90 עם עבודתו של Brent (1993) עבור אנגלית, והתפתח במהלך שני העשורים האחרונים תחת גישות שונות.

תהליך הלמידה של מסגרות ההצרכה אצל Brent נעשה בעזרתו של קורפוס שאינו מנותח או מתויג כלל. מכיוון שהקורפוס אינו מתויג משתמש Brent בתכונות השפה האנגלית וקובע את אפיונם של הפועל ומסגרת ההצרכה על סמך רמזים מורפו-סינטקטיים. מסגרת ההצרכה בעבודתו מוגבלת לתבניות מסוימות בלבד. המסגרות אותן בוחן Brent הן בנות ארגומנט אחד או שניים והארגומנטים יכולים להיות צירופים שמניים, שמות פועל או פסוקיות. תבניות מסגרות ההצרכה חייבות להופיע מיד לאחר הפועל. בעזרת הרמזים וההגבלות הללו אוסף Brent מידע סטטיסטי על היקרויות הפועל ומסגרות ההצרכה, וההחלטה על טיב הקשר ביניהם נעשית על סמך בדיקת השערות, עם המדד Binomial Hypothesis Test (BHT) וסף קבלה של 0.05. Brent בחן את שיטתו עבור 193 פעלים שונים, עבורם נבדקו באופן ידני זוגות הפועל-מסגרת הצרכה שהתקבלו. תוצאותיו מראות דיוק (precision) של 96% ואחזור (recall) של 60%. שיטתו של Brent אומצה גם עבור שפות נוספות, כדוגמת הונגרית (Simon et al. ,2010).

היתרון בשימוש בקורפוס שאינו מתויג הוא בכך שאין צורך להשתמש בכלים נוספים (שלא תמיד קיימים עבור שפות שונות) אלא רק בתכונות השפה. ההגבלות אותן מציב Brent על מסגרות ההצרכה שהוא בוחן ועל מיקומן ביחס לפועל, והצלחתו בזיהוי מסגרות ההצרכה של הפועל, מראות כי אין צורך בהכרח לנסות ולפתח שיטה כוללת ורחבה, אלא גם הצטמצמות לכדי זיהוי חלקי של מסגרות הצרכה מסוימות במיקום מסוים יכולה להניב תוצאות טובות. על בסיס זה פותחה שיטתנו לזיהוי מסגרות ההצרכה בעברית. השימוש במדדים סטטיסטיים שונים לבחינת הקשר בין הפועל למסגרת ההצרכה אף היא טכניקה אותה ניסינו ליישם במחקרנו, תוך הרחבת המדדים הסטטיסטיים.

Brent עצמו היה מודע למוגבלות של שיטתו אף עבור אנגלית, ובשל כך מספר התבניות של מסגרות ההצרכה שבחן הוא קטן. רק 6 מסגרות הצרכה שונות נבחנו, והן נבחרו משום שהיו קלות יותר לזיהוי, ועובדת הופעתם עם הפועל רמזה בהכרח לקשר ביניהם. בשל זאת, למשל, לא בחן Brent מסגרות הצרכה הכוללות מילות יחס, שכן, למרות שהוא מציין שקל לאתר את מילות היחס בקורפוס, הרי שבטכניקה שלו קיים קושי להבחין בין קשר של הצרכה או קשר תיאורי, כדוגמת מילות היחס המציינות מקום או זמן. בדומה ל-Brent אף אנו בחרנו להגביל את הארגומנטים של מסגרות ההצרכה של הפועל לקבוצה קטנה אפשרית, לאחר שבחנו את הופעת מסגרות ההצרכה בבנק העצים (פרק 4). כמו כן, זיהויים בקורפוס נעשה הן על פי הניתוח המורפולוגי (כדוגמת מילות היחס), והן על פי רמזים תחביריים (כדוגמת המילית ′ש′). כמו כן אימצנו אף אנו את הגישה לחיפוש מסגרת ההצרכה מיד לאחר הפועל (כאשר הלמידה נעשית מתוך קורפוס מתויג מורפולוגית בלבד, פרק 5). למרות שטבעי לעשות כן באנגלית, אך לא בעברית בה סדר המילים הוא חופשי יחסית ומסגרת ההצרכה יכולה להופיע גם לפני הפועל, התבוננות מעמיקה במסגרות ההצרכה בבנק העצים הובילה אותנו למסקנה כי ניתן לעשות זאת, ללא פגיעה רחבה בתהליך הלמידה (ראה סעיף 4.1 והדיון בעקבותיו בסעיף 5.1).

לאור ההתפתחות במשאבים לעיבוד שפות טבעיות, התפתחה גישה אחרת בעשור האחרון המשתמשת בכמה שיותר מידע על הטקסט על מנת ללמוד את מסגרות ההצרכה. בגישה זו נעשה שימוש בבנק עצים גדול המכיל, נוסף על התיוג של חלקי הדיבר, גם את הניתוח התחבירי והקשרים התחביריים במשפט. השימוש בניתוח התחבירי משפר את הדיוק של תהליך הלמידה, מכיוון שידוע עבור כל פועל מהן המילים הקשורות אליו, והתהליך בוחר מתוכן את מסגרת ההצרכה. כך נעשה בעבודתם של Sarkar ו-Zeman (2000) הלומדים את מסגרות הצרכה עבור הפועל בשפה הצ′כית. למידת המסגרות נעשתה בעזרת ה-Prague Dependency TreeBank, PDT, בנק עצים לשפה הצ′כית שנבנה ידנית. בניגוד לאנגלית, סדר המילים בצ′כית חופשי, ובשל כך למשלימי הפועל אין מקום קבוע ביחס לפועל. עקב כך לא ניתן להגדיר מראש תבניות למסגרת ההשלמה, כמו שקבע Brent, ולחפש אותן בקורפוס מיד לאחר הפועל. כמו כן, בשפה הצ′כית קיימות יחסות, ובשל כך מסגרת ההצרכה מורכבת יותר מאשר באנגלית. לדוגמה, עבור צירופי היחס המהווים ארגומנט לפועל, מלבד הדרישה למילת היחס המסוימת, נוספת הדרישה ליחסה שעל שם העצם לקבל. תהליך הלמידה של מסגרות ההצרכה נעשה בצורה איטרטיבית. בתחילת התהליך כל מופע של פועל בבנק העצים מגדיר מסגרת הצרכה פוטנציאלית רחבה הכוללת את כל האיברים הנמצאים בקשר עם הפועל בעץ הניתוח. בין איברים אלו מופיעים בערבוביה גם משלימים מוצרכים וגם תיאורים, ומטרת האלגוריתם היא להבחין ביניהם ולזהות את המשלימים. בכל שלב נבחן הקשר בין הפועל למסגרת בעזרת מבחנים סטטיסטיים שונים (Log Likelihood ratio, *t*-score, BHT), ואם המסגרת לא התקבלה תחת המבחנים השונים, מצטמצמת המסגרת באיבר אחד (שבחירתו אקראית), והמסגרת הקטנה יותר נבדקת שוב, עד להתייצבות. תוצאת עבודתם מראה שיפור בדיוק (precision) מ-55% ל-88%, לעומת בסיס (baseline) המתייג את כל המילים הקשורות לפועל

כתיאורים, ולא כמשלימים, ולכן מגדיר מראש את מסגרת ההצרכה של כל פועל כריקה. שימוש בבנק עצים באופן דומה נעשה עבור שפות שונות בהן סדר המילים הוא חופשי יחסית כדוגמת איטלקית (Ienco et al., 2008), וערבית (Attia et al., 2011).

עבור עברית, אמנם קיים בנק עצים, אך היקפו קטן יחסית (כ-6,500 משפטים, לעומת, לדוגמה, ה-PDT בו משתמשים Sarkar ו-Zeman הכולל 115,844 משפטים) ולכן הפעלת השיטה על עברית, בשל מספר הדוגמאות הקטן ובנוסף, בשל אי ההבחנה ברוב המקרים בין משלים מוצרך לתיאור (ראה סעיף 3.2), תיתן תוצאות לא מספקות. למרות זאת, המבחנים הסטטיסטיים אותם הם מבצעים לזיהוי הקשר בין הפועל למסגרת ניתנים לשחזור עבור מסגרות ההצרכה המצומצמות שאנו מחפשים בטקסט, ואכן בחרנו בעבודתנו להשתמש בהם לאור התוצאות הטובות שנתנו עבור השפות השונות.

לעומת הגישה האחרונה, המשתמשת לצורכי למידה בניתוח משפטים שנעשה ידנית, ניתן ללמוד גם ממשפטים שנותחו בעזרת מנתח אוטומטי (Briscoe & Carroll, 1997). עבודתם של Briscoe ו-Carroll הינה העבודה המרכזית שנעשתה בתחום ללמידת מסגרות הצרכה, ושימשה השראה לעבודות רבות כדוגמת Li ו-Brew (2005), Chesley ו-Salmon-Alt (2006) ואחרים. למידת מסגרות ההצרכה בגישה זו נעשית ראשית על ידי קביעת קבוצת מסגרות ההצרכה היכולות להופיע עם פעלים כלשהם (במקרה זה, המסגרות נלקחו מתוך לקסיקונים פועליים קיימים שנכתבו ידנית). לאחר קביעת מסגרות ההצרכה האפשריות, מנסים לאתר את אותן המסגרות הפוטנציאליות בסביבת כל מופע של הפועל, ולהעריך את הקשר ביניהם על סמך מדדים סטטיסטיים. המדד בו השתמשו Briscoe ו-Carroll הוא BHT (Binomial Hypothesis Test) כבעבודתו של Brent, אך בהמשך בוצעו הרחבות לעבודה זו על ידי שילוב מדדים סטטיסטיים נוספים (Korhonen, 2000). עבודתה של Korhonen משווה בין שלושה מדדים סטטיסטיים שונים: BHT (Binomial Hypothesis Test), LLR (Log Likelihood Ratio) ו-raw frequency. תוצאות העבודה הראו כי שימוש ב-raw frequency נתן את התוצאות הטובות ביותר, עם F-score של 65.2 (לעומת 53.3 ו-45.1 ב-BHT ו-LLR, בהתאמה). שיטה זו ללמידה נראית הטבעית ביותר ללמידת מסגרות הצרכה עבור כמות גדולה של טקסט, מכיוון שקיימת האפשרות לצמצם את החיפוש בתוך כל משפט לקבוצת המילים הקשורות לפועל בלבד. עבור עברית, בה סדר המילים במשפט חופשי יחסית, קיים יתרון נוסף לשיטה זו. בעזרת הניתוח התחבירי ניתן לזהות גם את חלקי מסגרת ההצרכה הפזורים במשפט כולו, לפני הפועל ואחריו, בצמידות לו ובמרחק מה ואין חובה להצטמצם לאזור הסמוך לפועל בלבד.

עבודה זו של Korhonen הורחבה לכדי יצירת לקסיקון של מסגרות הצרכה עבור אנגלית, VALEX (Korhone et al., 2006), המכיל בנוסף על מסגרות ההצרכה של הפועל גם מידע על ההסתברויות

שלהם. תוצאות עבודתם הראו F-score של 87.3. עבודתם של Korhonen ואחרים היוותה עבורנו מוטיבציה ליצירת לקסיקון פועל לעברית.

באותו אופן נעשתה עבודה ליצירת לקסיקון פועלי עבור צרפתית (LexSchem) (Messiant et al., 2008). Messiant ואחרים השתמשו בשיטה של Briscoe ו-Carroll ללמידת מסגרות ההצרכה, אך בניגוד ל-Korhonen ואחרים (2006) אין בעבודה שימוש בידע מוקדם על מסגרות ההצרכה, אלא מתבצעת למידה של המסגרות מתוך הנתונים. למידת המסגרות נעשית באופן דומה לזו של Sarkar ו-Zeman (2000). לכל פועל נלמדות מסגרות הצרכה היכולות להתאים לו, ואז מושמטות המסגרות הלא-סבירות. המדד בו הם משתמשים להכרעה הוא ה-raw frequency. מכיוון שבעברית אין כמעט לקסיקוני פועל, וגם אלו הקיימים אינם מקיפים דיים, גישה זו ללמידה יותר מתאימה לעברית מאשר גישתם של Korhonen ואחרים. בעבודתנו אנו מנסים ליישם גישה זו על עברית, תוך שימוש במנתח תחבירי שפותח לאחרונה (Goldberg, 2011). בניגוד ל-VALEX ו-LexSchem המכילים מסגרות הצרכה מלאות עבור הפעלים, הלקסיקון שלנו לא מכיל את מסגרת ההצרכה המלאה של הפועל, אלא רק מסגרות הצרכה חלקיות, המכילות ארגומנט יחיד בלבד, זאת בשל ההגבלות שהצבנו בתהליך הלמידה (ראה סעיף 5.1).

מאגר המידע היחיד עד כה על מסגרות ההצרכה של הפועל בעברית הוא מילון הפועל של שטרן (שטרן, 1994), לקסיקון פועלי שנכתב בצורה ידנית הכולל 833 פעלים שונים ו-1,430 מסגרות הצרכה שונות. מסגרות ההצרכה נלמדו מתוך קורפוס המכיל קטעי שיחה מתומללים, חיבורי תלמידים, טקסטים עיתונאיים, טקסטים ספרותיים וכן מתוך רשימות משלימים של פעלים שונים שהתקבלו מדוברי השפה. שטרן עושה הבחנה במשמעות של מסגרות שונות, וכולל בנוסף למבנה התחבירי של המסגרת גם את הגבלות הסמנטיות על כל ארגומנט. כמו כן, לכל מסגרת הצרכה מובאת עדות מהקורפוס בו השתמש על מנת להמחיש את המימושים השונים. למרות הניסיון ליצור לקסיקון מקיף, הרי שברור כי כל מילון ידני הוא מוגבל בהיקפו (לשם השוואה - מספר הפעלים הנמצא בלקסיקון מיל"ה הוא 4,804) וכל הרחבה של המילון תדרוש עבודה סיזיפית. כמו כן, מסגרת ההצרכה של הפועל יכולה להיות שונה ממשלב למשלב (Roland & Jurafsky, 1998) ולכן קיימת חשיבות ליצירת לקסיקון פועלי הנבנה על סמך המשלב של הקורפורה הקיימים לעברית, שיוכל להוות עוד נדבך ביצירת חבילת כלים שלמה עבור עברית. מלבד זאת, על מנת שיהיה ניתן להשתמש בלקסיקון לצרכים חישוביים, יש צורך גם בנתונים הסטטיסטיים על היקרויותיהם, לדוגמה, כאשר קיימות מספר מסגרות הצרכה לאותו פועל, אך יש עדיפות למסגרת זו או אחרת. העבודה הנוכחית פותרת את הליקויים הללו.

פרק 3

מתודולוגיית המחקר

בפרק זה נציג את מתודולוגיית המחקר בה אנו משתמשים (סעיף 3.1), את הכלים העומדים לרשותנו (סעיפים 3.1 ו-3.2) ואת השיטה להכרעת הקשר בין פועל למשלים (סעיף 3.3).

3.1 מחקר מבוסס ומונע קורפוס

מחקר מבוסס קורפוס הוא מחקר המשתמש ביכולת לעבד כמויות מידע גדולות, ומתבסס על העדויות העולות וצפות בקורפוס, על מנת ללמוד ולחקור תופעות לשוניות שונות.

עבודת המחקר נעשתה בעזרת ארבעה קורפוסים שונים של מיל"ה, מרכז הידע לתקשוב בשפה העברית (Itai & Wintner, 2008), המכילים טקסטים בעברית מודרנית. הקורפוסים מכילים בעיקר כתבות חדשותיות ומאמרים (הארץ, ערוץ 7, דה מרקר), ופרוטוקולי דיונים (הכנסת). נתונים סטטיסטיים על הקורפורה מופיעים בטבלה 3.1. הקורפורה מכילים 4,358 צורות בסיס של פעלים, ו-1,818,808 מופעי פעלים שונים.

שם הקורפוס	מספר תבניות	מספר תמניות	מספר צורות בסיס של פעלים	מספר מופעי הפעלים
הארץ	305,545	11,097,790	4,230	921,962
ערוץ 7	323,943	15,107,618	3,502	211,882
דה מרקר	62,216	692,919	2,729	55,972
הכנסת	204,967	15,066,731	3,766	628,992

טבלה 3.1 : נתונים על הקורפורה

Table 3.1: Corpora Data

3.1.1 עיבוד מוקדם של הקורפוס

הטקסטים בקורפורה עברו ניתוח מורפולוגי בעזרת המנתח המורפולוגי של מיל"ה (Itai & Wintner, 2008). המנתח המורפולוגי מפיק לכל תמנית בקלט את כל הניתוחים המורפולוגיים האפשריים, כאשר בין התכונות המורפולוגיות נכלל גם חלק הדיבר. מלבד הצגת כל הניתוחים

המורפולוגיים, קיים מתייג הקובע את חלק הדיבר הנכון בהקשר. עבור פעלים, ניתוחים שונים המתאימים לדרישות המתייג, כגון במין ובמספר, אך נבדלים בבניין, נחשבים כאפשריים, ללא בחירה בין הניתוחים השונים. כך לדוגמה עבור המשפט:

(א) לפני המשחק לא **חשבתי** על ניצחון או הפסד. (הארץ, 26.12.11)

יופיעו שני ניתוחים אפשריים לפועל **חשב**: האחד של חָשַׁב, בבניין פעל והשני של חִשֵּׁב, בבניין פיעל (איור 3.1). לניתוחים המורפולוגיים האפשריים ניתן ציון שונה מ-0, שהוא פונקציה רק של מספר הניתוחים האפשריים, כך שכל ניתוח אפשרי מקבל את אותו הציון.

```
<token id="4" surface="חשבתי">
    <analysis id="1" score="0.5">
        <base dottedLexiconItem="חָשַׁב" lexiconItem="חשב"
         lexiconPointer="12883" transliteratedLexiconItem="xeb">
            <verb binyan="Pa'al" gender="masculine and feminine"
             number="singular" person="1" register="formal" root="xeb"
             spelling="standard" tense="past"/>
        </base>
    </analysis>
    <analysis id="2" score="0.5">
        <base dottedLexiconItem="חִשֵּׁב" lexiconItem="חישב"
         lexiconPointer="2090" transliteratedLexiconItem="xieb">
            <verb binyan="Pi'el" gender="masculine and feminine"
             number="singular" person="1" register="formal" root="xeb"
             spelling="standard" tense="past"/>
        </base>
    </analysis>
</token>
```

איור 3.1 : ניתוח מורפולוגי של הפועל "חשבתי"

Figure 3.1 : Morphological analysis of '*xšbti*'

מלבד השימוש במנתח מורפולוגי השתמשנו בכלי שפותח במיל"ה על מנת לעדן את הציונים של הניתוחים המורפולוגיים לפעלים. עידון זה נעשה על סמך תפוצת בניינים, גופים וזמנים בקורפורה, ועל ידי העידון בציונים ניתן להעריך בצורה נכונה יותר את הניתוח הסביר ביותר. כך עבור משפט (א) לעיל הניתוח של הפועל **חשבתי** כפועל בבניין פעל מקבל ניקוד גבוה יותר על פני הניתוח השני, כפועל בבניין פיעל (איור 3.2).

```
<token id="4" surface="חשבתי">
    <analysis id="1" score="0.6675027978234862">
        <base dottedLexiconItem="חָשַׁב" lexiconItem="חשב"
         lexiconPointer="12883" transliteratedLexiconItem="xeb">
            <verb binyan="Pa'al" gender="masculine and feminine"
             number="singular" person="1" register="formal" root="xeb"
             spelling="standard" tense="past"/>
        </base>
    </analysis>
    <analysis id="2" score="0.33249720217651363">
        <base dottedLexiconItem="חִשֵּׁב" lexiconItem="חישב"
         lexiconPointer="2090" transliteratedLexiconItem="xieb">
            <verb binyan="Pi'el" gender="masculine and feminine"
             number="singular" person="1" register="formal" root="xeb"
             spelling="standard" tense="past"/>
        </base>
    </analysis>
</token>
```

איור 3.2 : ניתוח מורפולוגי של הפועל "חשבתי" עם ציון מעודן

Figure 3.2 : Morphological analysis of '*xšbti*' with refined score

3.2 בנק העצים

בנק העצים (Sima'an et al., 2001) הוא מאגר של כ-6,500 משפטים מנותחים מהמשלב העיתונאי, הלקוחים מתוך קורפוס הארץ. המשפטים המופיעים בבנק העצים עברו ראשית ניתוח מורפולוגי. לאחר הניתוח המורפולוגי עברו המשפטים ניתוח תחבירי ידני, ויוצגו במבנה של phrase-structure trees.

בניתוח התחבירי מצוינים, בין היתר, גם הקשרים בין הפועל למשלימיו, ועל כן מאפשר בנק העצים למידה של תכונות מסגרות ההצרכה. שלושה סוגי קשרים בין פועל למשלימו מסומנים בניתוח: קשר COM (complement), המהווה קשר של הצרכה, קשר DEP (dependency) שהוא קשר חלש יותר בין הפועל למשלימו, ובדרך כלל ניתן תיוג זה לקשר בין פועל לתיאורו, וקשר OBJ (object), בין פועל למושאו הישיר. מכיוון שההבחנה בין משלים מוצרך למשלים תיאורי הינה קשה אף לדוברים, הרי שההנחייה שניתנה למתייגים היא כי במקרה של ספק, יש לתאר את הקשר כ-DEP, ולכן ישנם מקרים רבים בהם קשר בין פועל למשלימו מתויג כך.

מתוך 6,500 משפטי בנק העצים אנו השתמשנו רק ב-3434 משפטים. קבוצה זו של משפטים כוללת אך ורק משפטים שאינם מכילים ציטוטים בתוכם, מכיוון שהקשרים בין חלקי המשפט שבתוך ומחוץ לציטוט כפי שהם מוצגים בבנק העצים הם סבוכים יותר, ולא תמיד משקפים נאמנה את התנהגות הפועל.

בניתוח בנק העצים נעזרנו בגרסת ה-Hebrew Dependency Treebank שפותחה על סמך גרסת ה- Hebrew Constituency Treebank של מיל"ה[1] (Goldberg, 2011). השימוש בגרסה זו של בנק העצים, ולא בגרסה המקורית, נעשה מכיוון שלצרכינו הספיק מבנה שטוח של הניתוח התחבירי, והייצוג ב-Hebrew Dependency Treebank התאים לדרישתנו, והיה קל יותר לניתוח.

3.3 בדיקת השערות (Hypothesis Testing)

על מנת לקבוע אם היקרות של פועל ומשלים יחדיו מצביעה על קשר של הצרכה או שזהו מופע מקרי, השתמשנו בבדיקת השערות (hypothesis testing) (Fisher, 1925). השערת האפס במקרה של למידת מסגרות ההצרכה היא כי אין קשר בין הפועל למשלימו, וההכרעה אם לקבל או לדחות את השערת האפס נעשית על ידי קביעת סף קבלה, הנקבע אמפירית (להסבר על אופן קביעת סף הקבלה ראה סעיף 5.3).

ארבעת המבחנים בהם השתמשנו ללמידת מסגרות ההצרכה הם ה-raw frequency של המשלים בהינתן הפועל (RF), log-likelihood, *t*-test ו-PMI. שלושת המבחנים האחרונים משמשים רבות לזיהוי קולוקציות, ובשל העובדה שיש בקשר ההצרכה דמיון לקולוקציה, משתמשים בהם גם בתחום זה (Sarkar & Zeman, 2000 ; Korhonen, 2002).

להגדרת המבחנים דרושות מספר הגדרות מקדימות. יהיו f מסגרת הצרכה המורכבת ממשלים אחד, ו-v פועל בצורת הבסיס שלו. על מנת לחשב את המדדים הסטטיסטיים נדרשים ארבעה ערכים אותם אומדים מתוך הקורפוס הנבדק, כאשר מופע של פועל בעל צורת בסיס v כולל כל נטיותיו בזמן, גוף ומין:

1. $n_{v,f}$ - מספר מופעי הפועל v עם המסגרת f
2. n_v - מספר מופעי הפועל v
3. $n_{\neg v,f}$ - מספר מופעי פועל כלשהו שאינו v עם המסגרת f
4. $n_{\neg v}$ - מספר מופעי כלל הפעלים השונים מ-v

ההגדרה של מופע פועל v עם מסגרת f תלויה בשיטת הלמידה. כשמסגרות ההצרכה נלמדות מתוך קורפוס מתויג מורפולוגית, מופע כזה הוא מופע בו המסגרת f מופיעה מיד לאחר הפועל v (הסבר על הסיבות לבחירת מסגרת הצרכה בת משלים אחד, ומיקומה ביחס לפועל, מובא בפרק 5). מאידך, בלמידת מסגרות הצרכה מתוך קורפוס מנותח תחבירית, מופע של פועל v עם מסגרת f נחשב ככזה אם קיים קשר של השלמה בין המסגרת (בת משלים אחד בלבד) לפועל (ראה פרק 7).

---

[1] תודתנו ליואב גולדברג על האפשרות להשתמש בגרסתו לצורך מחקרנו.

על סמך ערכים אלו נאמדות ההסתברויות הבאות (על פי maximum-likelihood estimation):

$$
\begin{aligned}
p_{f|v} &= \frac{n_{v,f}}{n_v} \\
p_{f|\neg v} &= \frac{n_{\neg v,f}}{n_{\neg v}} \\
p_f &= \frac{n_{v,f} + n_{\neg v,f}}{n_v + n_{\neg v}}
\end{aligned}
\tag{3.1}
$$

מכיוון שאין באפשרותנו לדעת אילו מופעים של המשלימים, כדוגמת מילות היחס או שמות הפועל, הם מופעים של מסגרת הצרכה, אנו מתייחסים רק לאותם משלימים המופיעים מיד לאחר הפועל, ומעריכים את ההסתברות של מסגרת ההצרכה, *p(f)*, להיות *p(f|any verb)*.

3.3.1 מדד Log Likelihood Ratio (LLR)

log-likelihood הוא מדד המאפשר להחליט אילו מבין ההשערות, השערת האפס או ההשערה האלטרנטיבית (H_1), היא סבירה יותר. המדד מתבסס על המאורע בו מתוך n_v מופעי הפועל v, $n_{v,f}$ מופעים הם עם מסגרת ההצרכה f, וכן מתוך $n_{\neg v}$ מופעי הפעלים השונים מ-v, $n_{\neg v,f}$ הם מופעים עם מסגרת ההצרכה f. המדד הוא היחס בין ההסתברות למאורע תחת ההשערה האלטרנטיבית לבין ההסתברות למאורע תחת השערת האפס, ותחת הנחת התפלגות בינומית, הוא מחושב באופן הבא:

$$
\begin{aligned}
LLR(v,f) = 2[&\log L\left(p_{f|v}, n_{v,f}, n_v\right) + \log L\left(p_{f|\neg v}, n_{\neg v,f}, n_{\neg v}\right) \\
&- \log L\left(p_f, n_{v,f}, n_v\right) - \log L\left(p_f, n_{\neg v,f}, n_{\neg v}\right)]
\end{aligned}
\tag{3.2}
$$

כאשר לכל p, k, n,

$$
\log L\left(p,k,n\right) = k \times \log p + \left(n-k\right) \times \log\left(1-p\right) \tag{3.3}
$$

השימוש ב-LLR לזיהוי קולוקציות הוצע על ידי Dunning (1993) והוא מציין כי התוצאות שהשיג תחת מבחן זה היו טובות יותר בהשוואה לשיטות שהיו רווחות באותה תקופה. היתרון של המבחן לשיטתו של Dunning הוא בכך שהוא אינו דורש כמות טקסט גדולה מאוד, ויכול לעבוד גם עם פחות טקסט מאשר המבחנים הסטטיסטיים שהיו נהוגים עד אז, כדוגמת מבחן χ^2. כמו כן המבחן יכול גם להציף תופעות שהן נדירות יחסית בטקסט. Dunning השתמש ב-LLR על מנת לזהות קולוקציות בנות שתי מילים בטקסט, והראה כי למבחן יתרון על מבחן χ^2.

בזיהוי מסגרות ההצרכה של הפועל נפוץ השימוש ב-LLR (Messiant et al., 2008; Korhonen, 2002; Korhonen et al., 2000; Sarkar & Zeman, 2000), אך ברוב המחקרים התוצאות המושגות על ידו אינן הטובות ביותר, ומדדים אחרים (כדוגמת BHT, binomial hypothesis test, בו השתמשו Briscoe ו-Carroll (1997)) מספקים תוצאות טובות יותר.

3.3.2 מדד *t*-score

t-score הוא מדד המאפשר לכמת את הסבירות לקבלת מדגם בעל התוחלת והשונות שהתקבלו, תחת ההנחה שהמדגם מתנהג על פי התפלגות נורמלית, בעלת תוחלת μ. המדד הוצע לראשונה לזיהוי קולוקציות על ידי Church, Gale, Hanks ו- Hindle (1991), ומאז נפוץ מאוד לזיהוי קולוקציות.

בעבודתם לזיהוי מסגרות הצרכה לפועל בשפה הצ׳כית השתמשו Sarkar ו-Zeman (2000) בוריאציה של מבחן זה לראשונה. המבחן בו השתמשו נגזר ממדד ה-*t*-score המקורי, בשינוי מה, והוא הפך לאחד המדדים הנפוצים בזיהוי מסגרות הצרכה (Korhonen, 2002). כאשר משתמשים במבחן לזיהוי מסגרות הצרכה, ערכו של ה-*t*-score משמש לקבוע את מידת הקשר בין הפועל למסגרת ההצרכה, והוא מחושב באופן הבא (Sarkar & Zeman, 2000):

$$T(v,f)=\frac{p_{f|v}-p_{f|\neg v}}{\sqrt{\sigma^2\left(n_v,p_{f|v}\right)+\sigma^2\left(n_{\neg v},p_{f|\neg v}\right)}} \tag{3.4}$$

כאשר לכל n ו-p השונות היא:

$$\sigma^2(n,p)=np(1-p) \tag{3.5}$$

החסם על מדד ה-*t*-score להכרעת הקשר בין הפועל למסגרת ההצרכה הינו חיובי. מדד ה-*t*-score בוחן האם מסגרת ההצרכה מעדיפה את הפועל הנוכחי על פני שאר הפעלים, על ידי הביטוי $p_{f|v}-p_{f|\neg v}$, כאשר ההנחה היא כי אם מסגרת ההצרכה קשורה לפועל, הרי ההסתברות להופעתה בהינתן הפועל, $p_{f|v}$, תהיה גבוהה מההסתברות להופעתה עם פועל אחר, $p_{f|\neg v}$. בשל כך אם קיים קשר בין מסגרת ההצרכה לפועל, אז ערך ה-*t*-score יהיה חיובי.

3.3.3 מדד Pointwise Mutual Information (PMI)

PMI הינו מדד סטטיסטי הבוחן קשר בין שתי מילים ונותן יכולת לאמוד האם הופעתן יחד היא מקרית או לא. Church ו-Hanks (1990) היו הראשונים להשתמש במדד ה-PMI לזיהוי קולוקציות

(collocations), ולזיהוי קשר של הצרכה בין פועל לשם פועל, תוך הבחנה בין שמות הפועל למופעי פועל עם מילת היחס to. במקרה של למידת מסגרות הצרכה הקשר בין הפועל v למסגרת ההצרכה f הוא הנבחן. אם היקרותם של הפועל והמסגרת יחד היא מקרית והופעת מסגרת ההצרכה אינה תלויה בפועל אזי מתקיים כי $p(v,f) = p(v)p(f)$. המדד מחושב באופן הבא:

$$I(v, f) = \log \frac{p(v, f)}{p(v)\, p(f)} \tag{3.6}$$

לא ידוע לנו על ניסיון לזיהוי מסגרות ההצרכה של הפועל תחת מדד זה, אך מכיוון שהשימוש בו לזיהוי קולוקציות נותן תוצאות טובות, אנו משערים כי הבחירה בו תוכל להועיל בזיהוי הקשר בין פועל למסגרת.

מדד ה-PMI בוחן אם הופעת מסגרת ההצרכה היא אקראית, או תלויה בפועל, זאת על ידי הקשר $p(v, f) > p(v) \cdot p(f)$ בין ההסתברויות המצביע על ייחודיות הזוג פועל-מסגרת הצרכה, על פני מופע אקראי, דהיינו $I(v, f) = \log \frac{p(v, f)}{p(v)\, p(f)} > 0$, ולכן אם קיים קשר בין מסגרת ההצרכה לפועל, ערך ה-PMI יהיה חיובי.

נשים לב, כי מדד ה-PMI ומדד ה-t-score מקבלים ערך חיובי בדיוק באותם תנאים (משוואה 3.7).

$$\begin{aligned}
&PMI : p(v, f) > p(v)p(f) \Leftrightarrow \frac{n_{v,f}}{N} > \frac{n_v}{N} \frac{n_{v,f} + n_{\neg v,f}}{n_v + n_{\neg v}} \Leftrightarrow \\
&\frac{n_{v,f}}{n_v} > \frac{n_{v,f} + n_{\neg v,f}}{n_v + n_{\neg v}} \Leftrightarrow n_{v,f}\, n_{\neg v} > n_{\neg v,f}\, n_v \Leftrightarrow \\
&\frac{n_{v,f}}{n_v} > \frac{n_{\neg v,f}}{n_{\neg v}} \Leftrightarrow p(f \mid v) > p(f \mid \neg v) : T - score
\end{aligned} \tag{3.7}$$

ההבדל בין המדדים הללו, אם כן, הוא מידת הזיקה שהם קובעים בין הפועל למסגרת ההצרכה.

פרק 4

מסגרת ההצרכה בבנק העצים: בחינת תופעות לשוניות

בנק העצים שתואר בסעיף 3.2 יכול לשמש מקור ללמידת תופעות לשוניות הקשורות למסגרת ההצרכה. למרות הבחינה הדקדקנית בתופעות הלשוניות שנעשתה על ידי בלשנים רבים, יש צורך גם בבחינה כמותית של התופעות. הדבר חיוני במיוחד כהקדמה למחקר סטטיסטי, כמחקרנו, הבא לנסות ולזהות את התופעה במרחב המילה הכתובה.

במסגרת בחינה זו נבדקו כ-3,434 משפטים מתוך בנק העצים, הכוללים 1,423 פעלים שונים ו-7,561 מופעי פועל.

כמו שציינו בסעיף 3.2 הקשרים הקיימים בבנק העצים בין הפועל למשלימו הם COM (complement), DEP (dependency) ו-OBJ (object). מכיוון שההנחייה שניתנה למנתחים היא כי במקרה של ספק, יש לתאר את הקשר כ-DEP, הרי שקיימים מקרים רבים של ניתוח לא מדויק דיו של הקשרים. בשל כך, בבחינה זו אנו בוחנים את כלל המשלימים המקיימים קשר כלשהו עם הפועל (OBJ, DEP ו-COM), אלא אם כן צוין אחרת.

4.1 מרחק המשלים מהפועל

התבוננות על מיקומי המשלימים ביחס לפועל מלמדת כי רוב מופעי המשלים נמצאים מיד לאחר הפועל (52.42% מהמשלימים). כמו כן רוב המשלימים נמצאים לאחר הפועל (87.46%, שהם 8,296 מהמופעים). בטבלה 4.1 מופיעה התפלגות מרחק המשלים מהפועל במילים, עבור מרחקים המופיעים מעל ל-1% בבנק העצים.

מרחק מהפועל	אחוז	מספר מופעים
< -5	4.70%	446
-5	1.33%	126

מרחק מהפועל	אחוז	מספר מופעים
-4	1.56%	148
-3	2.44%	231

מרחק מהפועל	אחוז	מספר מופעים
-2	2.40%	228
-1	0.12%	11
1	52.42%	4,973
2	7.22%	685
3	8.39%	796
4	6.11%	580
5	3.16%	300
6	2.68%	254
7	1.57%	149
8	1.43%	136
>8	4.46%	423

טבלה 4.1 : מרחק המשלימים מהפועל

Table 4.1 : Verb-complements distance

4.2 מילות היחס

בבנק העצים 129 מילות יחס שונות נמצאות ביחס כלשהו לפועל (5,983 מופעים שונים). מטבלה 4.2 עולה כי מילות היחס הנפוצות ביותר הן ב, ל, מ, על, עם, אל, והן מהוות כ-77.67% (4,647 מופעים) מכלל הופעות מילות היחס הקשורות לפועל, ולכן בחרנו להתמקד רק בהן (ראה פרק 5) .

מילת יחס	אחוז מופעים מכלל מילות היחס	מספר מופעים
ב	39.75%	2,378
ל	19.40%	1,161
מ	7.55%	452
על	7.54%	451
עם	2.27%	136
אל	1.15%	69
סכום	**77.67%**	**4,647**

טבלה 4.2 : התפלגות מילות היחס

Table 4.2 : Prepositions distribution

4.3 צירוף שמני

קיימים 2,207 מופעי צירוף שמני כמשלים לפועל, מתוכם כ-51.7% (1,141 מופעים) הינם מופעים שאינם מיודעים, ולכן לא מופיעה לפניהם המילית ׳את׳. זיהוי הצירוף השמני כמשלים רק על פי המילית ׳את׳ יגרום לכך שהמספרים לא יישקפו נאמנה את תכונת הפועל, מנגד, לא ניתן להחשיב את כל הצירופים השמניים המופיעים מיד לאחר הפועל כמשלימים, מכיוון שלאחר הפועל יכול להופיע גם נושא.

4.4 מופעי הנושא במשפט

בבנק העצים רק ל-3,692 ממופעי הפעלים יש נושא המקושר אליהם. ב-32.48% מהמופעים הנושא מופיע מיד לפני הפועל (במרחק 1-), וב-12.89% מהמופעים הנושא מופיע מיד אחרי הפועל. בטבלה 4.3 מופיעה התפלגות מרחק הנושא מהפועל, עבור מרחקים המופיעים מעל ל-1% בבנק העצים.

המרחק	אחוז	מספר
< -11	5.31%	196
-11	1.03%	38
-10	1.14%	42
-9	1.22%	45
-8	1.65%	61
-7	2.11%	78
-6	2.55%	94
-5	2.87%	106
-4	3.49%	129

המרחק	אחוז	מספר
-3	8.21%	303
-2	9.70%	358
-1	32.48%	1,199
1	12.89%	476
2	8.94%	330
3	2.52%	93
4	1.54%	57
> 4	2.36%	87

טבלה 4.3 : התפלגות מרחק הנושא מהפועל

Table 4.3 : Object-verb distance distribution

4.5 פסוקית

בחינה של 554 מופעי הפסוקיות הקשורות לפועל (טבלה 4.4) מראה כי אין העדפה מובהקת לאחת ממילות השעבוד.

מילת שעבוד	אחוז המופעים	מספר מופעים
כי	49.28%	273
ש	50.72%	281

טבלה 4.4 : התפלגות מילות השעבוד

Table 4.4 : Complementizer distribution

4.6 המשלימים

התפלגות המשלימים מראה כי רוב המשלימים לפועל הם מילות יחס (כ-63% מהמשלימים). מילות היחס הנבחנות: אל, ב, ל, מ, על ו-עם מהוות כ-49% מכלל המשלימים.

סוג משלים		אחוז מופעים	מספר מופעים
מילות יחס	מילות יחס נבחנות: אל, ב, ל, מ, על, עם	48.99%	4,647
	מילות יחס אחרות	14.09%	1,337
צירוף שמני		23.27%	2,207
שם פועל		6.58%	624
פסוקית		5.84%	554

טבלה 4.5 : התפלגויות המשלימים

Table 4.5 : Complements distribution

4.7 מספר משלימים במסגרת ההצרכה

כאשר בוחנים את מספר המשלימים במסגרות ההצרכה המופיעות בבנק העצים (טבלה 4.6), תוך התייחסות לקשר הרחב ביותר עם הפועל (כולל DEP), ניתן לראות כי ל-68.39% ממופעי הפעלים יש לכל היותר משלים אחד.

מספר משלימים	אחוז מופעים	מספר מופעים
0	14.14%	1,069
1	54.15%	4,094
2	25.99%	1,965
3	5.13%	388
4	0.58%	44
5	0.01%	1

טבלה 4.6 : התפלגות מספר המשלימים - ALL

Table 4.6 : Number of complements distribution - ALL

מכיוון שהקשר הרחב יכול לכלול גם משלימים שאינם מוצרכים לפועל, בדקנו אף את התפלגות מספר המשלימים לכלל הפעלים תחת קשר מחמיר יותר של הכולל רק את OBJ ו-COM (טבלה 4.7). במקרה זה ל-93.61% ממופעי הפעלים יש משלים אחד לכל היותר. כמו שציינו לעיל, קשר ה-COM הוא מחמיר יותר, וייתכן כי ישנם מופעים רבים של פעלים עבורם הקשר עם המשלימים המוצרכים תויג כקשר DEP, ועל כן האחוז הנכון נמצא בין שני הנתונים לעיל, אך בכל מקרה, אחוז גבוה ממופעי הפעלים אינו דורש יותר ממשלים אחד. בשל כך, בחרנו לבחון בעבודתנו מסגרות הצרכה חלקיות בנות משלים יחיד בלבד.

מספר משלימים	אחוז מופעים	מספר מופעים
0	37.85%	2862
1	55.76%	4216
2	6.34%	479
3	0.05%	4

טבלה 4.7 : התפלגות מספר המשלימים - COM

Table 4.7 : Number of complements distribution - COM

פרק 5

למידת מסגרות ההצרכה מתוך קורפוס מתויג מורפולוגית

כאמור בפרק 1.3, מסגרת ההצרכה היא בעלת חשיבות רבה ליישומים חישוביים שונים. למיטב ידיעתנו, טרם נעשתה אף עבודה למציאת מסגרות ההצרכה בעברית בצורה אוטומטית. מטרת מחקרנו היא לפתח שיטה ללמידת מסגרות ההצרכה של הפועל מתוך קורפוס, ובעקבותיה ליצור לקסיקון פועלי מקיף שיכיל עבור הפעלים את מסגרות ההצרכה שלהם, וכן מידע סטטיסטי על שכיחויות המסגרות השונות. במסגרת פרק זה נאפיין במדוייק את מסגרות ההצרכה אותן אנו בוחנים (סעיף 5.1) נתאר את השיטות בהן השתמשנו לצורך זיהוי מסגרות ההצרכה בקורפוס (סעיף 5.2) ואת אופן הערכת הקשר בין הפועל למסגרת ההצרכה (5.3). לבסוף נציג את תוצר התהליך - לקסיקון הפועל (סעיף 5.4)

חלקו הראשון של מחקרנו מתרכז בלמידה של מסגרת ההצרכה מתוך קורפוס מתויג מורפולוגית. בהמשך העבודה (פרק 7) נציג את גישתנו ללמידה של מסגרת ההצרכה מתוך קורפוס מנותח תחבירית.

5.1 איפיון מסגרת ההצרכה

הבעיה הראשונית העומדת בפנינו היא בעיית זיהוי המשלימים היכולים להוות חלק ממסגרות ההצרכה של הפועל במשפט.

מסגרת ההצרכה בעבודה זו הינה אוסף של צירופים שמניים, צירופי יחס, פסוקיות ושמות פועל הקשורים לפועל. אנו מגבילים את קבוצת מילות היחס שיכולות להופיע תחת צירופי היחס להיות מילות היחס: אל, ב, ל, מ, על ו-עם. הגבלה זו נעשית על סמך מילות היחס הנפוצות בעברית (4.1), ועל פי ההגבלה המקובלת בספרות (אזר, תשל"ב). כמו כן אנו מגבילים את הפסוקיות המשועבדות רק לכאלה המתחילות במיליות הקישור 'ש' ו-'כי' (4.5).

בקורפוס העובר תיוג מורפולוגי בלבד, מה שעומד לרשותנו לצורכי זיהוי מסגרות ההצרכה הוא תיוג חלקי הדיבר. בעזרת תיוג חלקי הדיבר אנו כמובן יכולים לזהות את הפעלים, את הצירופים השמניים,

צירופי היחס, הפסוקיות ושמות הפועל, אך לא לזהות האם יש ביניהם לפועל קשר, ואם כן - מהו. גישה פשטנית, בה מחשיבים את כל המילים המתויגות בתיוג המתאים להגדרת המשלים כמסגרת הצרכה, וממנה מסיקים על מסגרות ההצרכה השלמות של הפועל לא תניב תוצאות משביעות רצון. הסיבה לכך נעוצה, בין היתר, בעובדה כי למילים זהות, בעלות אותו חלק דיבר, יכולים להיות תפקידים תחביריים שונים. ניתן לראות זאת במשפטים הבאים:

א) חשבתי **על** הספר **ש**קראתי אתמול.

ב) חשבתי **ש**הספר **על** המדף.

בהם הפועל **חשבתי** מופיע, אך במשפט הראשון עם מסגרת ההצרכה **על** ואילו במשפט השני עם מסגרת ההצרכה **ש**. לקיחת **על** ו-**ש** יחדיו תוביל להסקת מסגרות הצרכה שאינן נכונות (**חשב** + על + ש). כמו כן, המשפטים יכולים להיות מורכבים ממספר פעלים, במיוחד במשלב העיתונאי המאפיין את הקורפוס שלנו, בו המשפטים נוטים להיות ארוכים ומורכבים. בשל כך, במשפט אחד מופיעות מספר מסגרות ההצרכה שונות, ויש להבחין ביניהן, ולשייך כל אחת מהן לפועל המתאים. על כן יש לפתח טכניקה לזיהוי המשלימים היכולים להיות חלק ממסגרת ההצרכה.

לפיכך, יצרנו מערכת אילוצים על הזוגות אותם אנו מזהים:

1. **סוג המשלים:** כמו שציינו המשלימים שאנו בוחנים הם צירופים שמניים, צירופי יחס בהם מילת היחס היא: אל, ב, ל, מ, על, עם, פסוקיות ושמות פועל בלבד.
2. **מיקום המשלים:** אנו מתייחסים רק לצירופים המופיעים מיד לאחר הפועל כמשלימים פוטנציאליים, ועקב זאת מסגרת ההצרכה שאנו בוחנים היא בת איבר אחד בלבד.

אמנם מסגרת ההצרכה יכולה להיות רחבה ובת יותר ממשלים אחד (סעיף 1.2), והמשלימים יכולים להופיע לפני ואחרי הפועל, ולאו דווקא בסמיכות אליו (זאת בשל סדר המילים החופשי יחסית במשפט בעברית (Shlonsky ,1997), אולם בחינה של הנתונים מבנק העצים (פרק 4) מראה כי הטכניקה בה נקטנו משקפת את המציאות התחבירית. כמו כן, בשל סדר המילים החופשי, עבור פעלים בעלי ערכיות תחבירית גבוהה אין העדפה למשלים זה או אחר בקרבה לפועל, ולכן ניתן יהיה לקבל את המסגרת התחבירית כולה על ידי שילוב המידע שאספנו עבור הפועל.

3. **הגבלה על הצירופים השמניים:** אנו מתייחסים כמשלים רק לצירופים שמנים המתחילים במילית 'את' או שגרעינם אינו מתאים במין או במספר לפועל.

העובדה כי קיימים מופעי משלימים שאינם מצוינים על ידי סמן היחסה 'את' מהווה בעיה כאשר רוצים למצוא באופן אוטומטי את משלימיו של הפועל (4.3). הבעיה העיקרית, הנובעת מסדר המילים החופשי של המשפט בעברית, היא כי מופע של צירוף שמני, גם כזה המציית להגבלה 2, ומופיע מיד לאחר

הפועל, יכול להיות בתפקיד נושאי (4.4), ולכאורה אין דרך להבדיל ביניהם. אולם שימוש בתכונות המורפולוגיות של הפועל והצירוף השמני יכול להיות לעזר בהבחנה בין מופע נושאי או מושאי. בעברית יש התאמה במין ובמספר בין גרעין הנושא והפועל, ועל כן הוספת קריטריון הבחנה של חוסר התאמה ימנע ממופעי צירוף שמני נושאי להילקח כמשלים.

בנוסף למסגרות ההצרכה שהצגנו לעיל, פועל יכול לדרוש מסגרת הצרכה שהיא ריקה, כלומר, לא דורש כלל משלימים, ויכול לעמוד בפני עצמו.

לסיכום, מסגרת ההצרכה שאנו בוחנים היא מסגרת בת ארגומנט אחד (לכל היותר), שהוא צירוף שמני, צירוף יחס עם מילת היחס: אל, ב, ל, מ, על, עם, פסוקית או שם פועל, המופיעה מיד לאחר הפועל או מסגרת ריקה.

5.2 איסוף מידע סטטיסטי על הפועל ומסגרות ההצרכה

תהליך למידת מסגרות ההצרכה של כל פועל מורכב משני חלקים. חלקו הראשון של התהליך הוא זיהוי של מסגרות הצרכה פוטנציאליות עבור כל פועל ואיסוף של מידע סטטיסטי על היקרויותיהם של מסגרות ההצרכה והפעלים, ואילו חלקו השני של התהליך נוגע להכרעה בדבר הקשר בין הפועל למסגרות ההצרכה. סעיף זה יעסוק בתהליך איסוף המידע הסטטיסטי.

בשל ההגבלות שתיארנו לעיל אנו מצטמצמים לבחינת עשר מסגרות הצרכה שונות: צירוף שמני, מילת יחס – אל, ב, ל, מ, על, עם, פסוקית, ושם פועל, אותן אנו מחפשים מיד לאחר הפועל ומסגרת ריקה. על מנת לקבוע אילו מבין עשר מסגרות ההצרכה מתאימות לפועל בצורת הבסיס שלו, v, אנו בוחנים את כל מופעיו של הפועל בקורפוס ואוספים נתונים סטטיסטיים על תפוצתם של הפועל והמסגרות השונות המופיעות מיד לאחריו. נציין כי כאשר אנו מונים מופעי פועל בתהליך זה, אנו מתייחסים לכל מופעיו של הפועל בקורפוס, על כל נטיותיו בזמן, גוף ומין. מכיוון שאנו מתרכזים באלמנט המופיע מיד לאחר הפועל, כאשר לא מופיעה אחת מתשע מסגרות ההצרכה מיד לאחריו אנו מחשיבים מופע זה של הפועל כמופע של פועל עם מסגרת הצרכה ריקה (אין הדבר רומז לכך שהפועל מופיע בסוף המשפט ואין דבר אחריו, אלא רק מתייחס לעובדה כי אף אחת ממסגרות ההצרכה הנבחנות לא מופיעה מיד לאחריו).

ארבעה ערכים חשובים להכרעה בדבר הקשר בין הפועל v למסגרת f ואותם אנו מחשבים בתהליך:

1. $n_{v,f}$ - מספר מופעי הפועל v עם המסגרת f בקורפוס (כלומר, מספר מופעי הפועל כשמיד לאחריו מילת יחס, מילת שעבוד, שם פועל או מסגרת ריקה, לפי המסגרת).
2. n_v - מספר מופעי הפועל v בקורפוס.

3. $n_{\neg v, f}$ - מספר מופעי פועל כלשהו שאינו v עם המסגרת f (כלומר, מספר מופעי פעלים שונים שאינם v, כשמיד לאחריהם מילת יחס, מילת שעבוד, שם פועל או מסגרת ריקה לפי המסגרת).
4. $n_{\neg v}$ - מספר מופעי כלל הפעלים השונים מ-v.

נעיר כי כמו שציינו בסעיף 3.1.1, הניתוח המורפולוגי אינו שלם, שכן למרות שאנו יודעים לזהות את הפעלים בקורפוס, איננו יודעים מהו הניתוח הנכון, אלא יש לפנינו אוסף ניתוחים אפשריים. ולכן במשפטים בהם יש עמימות לגבי הניתוח הנכון (איור 3.2), כמו במשפט:

(א) לפני המשחק לא **חשבתי על** ניצחון או הפסד. (הארץ, 26.12.11)

אנו צריכים לקחת בחשבון את שני הניתוחים האפשריים, ולמנותם כשני מופעים של שני פעלים שונים: האחד של הפועל חָשַב + על והשני חִשֵּב + על. מכיוון שמנייה של שני המופעים האלו כאילו הם שונים תשפיע באופן לא רצוי על למידת המסגרות, אנו מתחשבים בניקוד שניתן לכל ניתוח, כך שכל מניה מוכפלת בפקטור שהוא הניקוד הניתן לניתוח, ולכן במקרה שלנו התוספת ל-$n_{\text{על, חשב}}$ היא לא של 1, אלא של 0.667. באופן הזה נמנים אמנם גם צירופים שלא היינו רוצים לקבל, כמו "חישב על", אך בשל הניקוד הנמוך אותו הם מקבלים, וכן בשל העובדה שקיימים מופעים אחרים, נכונים של הפועל "חישב", בהם מופיעה מסגרת ההצרכה הנכונה, אנו צופים כי ניתוחים לא נכונים לא יצופו.

בתהליך איסוף המידע הסטטיסטי אספנו 20,829 זוגות של פועל בצורת הבסיס-מסגרת הצרכה המופיעה מיד אחריהם, ביניהם 3,393 צורות בסיס שונות של פעלים, ולכל זוג כזה חישבנו את ארבעת הערכים המוזכרים לעיל: $n_{v,f}$, n_v, $n_{\neg v, f}$ ו-$n_{\neg v}$.

5.3 הערכת הקשר בין הפועל למסגרת הצרכה

לאחר תהליך איסוף הזוגות של פעלים עם מסגרות ההצרכה המופיעות איתם והמידע הסטטיסטי, עלינו להכריע האם המשלים שהופיע בסמיכות לפועל הוא אכן חלק ממסגרת ההצרכה של הפועל. הכרעה כזו הכרחית שכן לא כל המסגרות המופיעות עם הפועל אכן ממלאות תפקיד זה. לדוגמה, ישנם משלימים הנפוצים מאוד בשפה, כדוגמת צירופי מקום או זמן הבאים לידי ביטוי במילות היחס **ב**, **מ**, **ל**, **אל**, שעובדת הופעתם לצד הפועל, אין בה בכדי לרמז על היותם חלק ממסגרת ההצרכה שלו. לצורך אומדן וכימות הקשר בין הפועל והמסגרת אנו משתמשים במדדים סטטיסטיים, ובעזרת בדיקת השערות מזהים את ההצרכה. המדדים הסטטיסטיים בהם השתמשנו הם: PMI (*Pointwise Mutual Information*), *LLR* (*Log Likelihood Ratio*), *t-score* ו-RF (*Raw Frequency*), שהם מדדים מקובלים לזיהוי קשר ייחודי בין אלמנטים שונים (סעיף 3.3).

את ההסתברויות אנו אומדים על פי maximum-likelihood estimation (5.1), ובעזרתן מחשבים את המדדים השונים (ראה סעיף 3.3 להסבר על אופן החישוב של כל מדד). כמו כן, אנו מעריכים את ההסתברות של מסגרת ההצרכה, *p(f)*, להיות *p(f|any verb)*.

$$
\begin{aligned}
p_{f|v} &= \frac{n_{v,f}}{n_v} \\
p_{f|\neg v} &= \frac{n_{\neg v,f}}{n_{\neg v}} \\
p_f &= \frac{n_{v,f} + n_{\neg v,f}}{n_v + n_{\neg v}}
\end{aligned}
\tag{5.1}
$$

לאחר חישוב המדדים השונים עבור זוגות פועל-מסגרת הצרכה יש צורך בסף קבלה אשר מעליו דוחים את השערת האפס (לא קיים קשר ייחודי בין הפועל למסגרת ההצרכה), ומקבלים את הזוג פועל-מסגרת הצרכה. למרות שלחלק מהמבחנים הסטטיסטיים קיימות טבלאות הקובעות את סף הקבלה, כתלות ברמת המובהקות שרוצים להשיג ומספר דרגות החופש, קיים ערך מוסף ללמידת הספים בצורה אמפירית על משימה הדומה למטרת העבודה, כך שספי הקבלה משקפים נאמנה את חוזק הקשר הנדרש לקביעת מסגרות ההצרכה. בחרנו לקבוע את הספים על סמך בדיקות שערכנו על חלקו של בנק העצים ששימש לפיתוח. חלק הפיתוח בו השתמשנו מהווה 20% מהמשפטים בבנק העצים (3.2), וכולל 1,057 משפטים בהם 676 צורות בסיס של פעלים ו-1,536 מופעי פועל על נטיותיו השונות.

קביעת סף הקבלה נערכה עבור כל מדד בנפרד. לצורכי הערכת סף הקבלה האופטימלי בחנו ערכי סף שונים בתחום הערכים האפשרי של כל מדד, ועבור כל אחד מהם הערכנו את מסגרות ההצרכה שהתקבלו אל מול מסגרות ההצרכה שהיו אמורות להתקבל על פי בנק העצים.

קבוצת הפעלים הנבחנים מכילה 410 פעלים בצורת הבסיס. כל הפעלים נבחנים עם 9 מסגרות הצרכה אפשריות (כל מסגרות ההצרכה האפשריות, פרט למסגרת ההצרכה הריקה), (סה"כ 3,690 זוגות פועל-מסגרת הצרכה שונים). הקביעה האם זוג פועל-מסגרת הצרכה מתקבל נעשית על סמך הציון שקיבל במדד. אם הציון הוא מעל סף הקבלה הנבדק, אזי הוא מתקבל. התוצאות מוערכות אל מול קבוצת זוגות פועל-מסגרת הצרכה בת 875 זוגות שונים, המתקבלת מבנק העצים. זוגות אלו מאופיינים בקשר COM (סעיף 3.2) המופיע ביניהם. כמו שציינו בפרק 3, תיוג קשר בין משלים לפועל כקשר COM נעשה בצורה מחמירה, ולכן קיימים מופעים רבים של פועל המופיעים עם משלים שהוא חלק ממסגרת ההצרכה ולמרות זאת הקשר ביניהם אינו מתויג כקשר COM אלא כקשר DEP. בשל כך מופעים רבים של פעלים, לכאורה, הם מופעים של פועל עם מסגרת הצרכה ריקה. מאידך, לקיחת כל המשלימים של הפועל הנמצאים עימו בקשר DEP יכולה להוביל להערכה מוטעית של משלימים שאינם

חלק ממסגרת ההצרכה. בשל כך בחרנו להשתמש רק בקשרי COM עם הפועל, ולא להשתמש במסגרות ההצרכה הריקות לשם הערכת ספי הקבלה.

הערכת ספי הקבלה השונים מתבססת על המונים הבאים:

- TP (true positive) - מספר הזוגות אשר התקבלו בקבוצה הנבדקת ונמצאים בקבוצה מבנק העצים
- TN (true negative) - מספר הזוגות אשר נדחו בקבוצה הנבדקת ואינם נמצאים בקבוצה מבנק העצים
- FP (false positive) - מספר הזוגות אשר התקבלו בקבוצה הנבדקת ואינם נמצאים בקבוצה מבנק העצים
- FN (false negative) - מספר הזוגות אשר לא התקבלו בקבוצה הנבדקת אך הם נמצאים בקבוצה מבנק העצים

בעזרת מונים אלו אנו מחשבים את הדיוק (precision), אחוז הזוגות הנכונים מכלל הזוגות שהתקבלו, והאחזור (recall), אחוז הזוגות הנכונים שהתקבלו מתוך כלל הזוגות שהיו אמורים להתקבל. כמקובל, המדד המשמש למיצוע בין שני המדדים הללו הוא ה- F-score - הממוצע ההרמוני של הדיוק והאחזור.

$$\begin{aligned} \text{precision} &= \frac{TP}{TP+FP} \\ \text{recall} &= \frac{TP}{TP+FN} \\ \text{F-score} &= 2\cdot\frac{\text{precision}\cdot\text{recall}}{\text{precision}+\text{recall}} \end{aligned} \tag{5.2}$$

לאחר חישוב מדד ה-F-score עבור ספי הקבלה השונים, נבחר סף הקבלה עבור המדד הנבדק ככזה הממקסם את מדד ה-F-score. ספי הקבלה שהתקבלו לאחר התהליך הנ"ל הם:

מדד	סף קבלה
LLR	544.17
PMI	0.12
RF	0.11
t-score	0.12

טבלה 5.1 : ספי הקבלה עבור המבחנים הסטטיסטיים השונים

Table 5.1 : Thresholds of the statistical tests

5.4 לקסיקון הפועל

לאחר תהליך הוצאת הזוגות מהקורפורה קיבלנו 20,829 זוגות שונים של פועל-מסגרת הצרכה, עבור 3,393 צורות בסיס של פעלים. עבור כל מדד נבנה לקסיקון פועל המכיל זוגות של פועל-מסגרת הצרכה שנלמדו בתהליך, וציונם תחת המדד היה מעל לסף הקבלה של המדד. באופן הזה התקבלו 4 לקסיקוני פועל שונים, השונים זה מזה בפעלים אותם הם כוללים, ובמסגרות ההצרכה המתאימות לכל פועל. נתונים על לקסיקוני הפועל השונים ניתן לראות בטבלה 5.2.

מדד	מספר פעלים בצורת הבסיס	מספר זוגות פועל-מסגרת הצרכה	מספר מסגרות ממוצע לפועל
LLR	413	706	1.71
PMI	3,393	7,139	2.10
RF	3,393	7,392	2.18
t-score	3,390	5,633	1.66

טבלה 5.2 : נתונים על מספר הזוגות פועל-מסגרת הצרכה שהתקבלו תחת המדדים השונים

Table 5.2 : Verb-frames couples Data

פרק 6

תוצאות והערכה

תוצאות התהליך שתיארנו בפרק 5 הינם ארבעה לקסיקוני פועל, שכל אחד מהם הוא תוצאה של הפעלת מדד סטטיסטי שונה. על מנת להעריך את איכות התוצאות שהתקבלו בחרנו שתי גישות שונות. הגישה הראשונה מתמקדת בקבוצת פעלים מצומצמת ובוחנת את התוצאות אל מול ניתוח ידני שנעשה עבורם (6.1). מכיוון שגישה זו, בשל העובדה שהניתוח נעשה באופן ידני, היא מוגבלת בהיקפה, בחרנו להעריך את איכות התוצאות שהתקבלו גם לא במישרין, אלא על ידי שימוש בתוצאות שקיבלנו לשיפור הביצועים של שתי משימות חישוביות: הכרעה בבעיית הצמדת צירוף היחס (6.3), ותרגום אוטומטי (6.4). כמו כן נקיים דיון על יכולות הזיהוי של מסגרת ההצרכה תחת המדדים הסטטיסטיים בהם השתמשנו, נסקור גורמים שונים שהשפיעו על התוצאות שהתקבלו ונבחן תופעות שונות שהופיעו בהן (6.2).

על מנת לאפשר את ההערכה לאורך פרק זה, קבענו בצורה ידנית עבור כל אחד מהפעלים המופיעים בפרק את קבוצת מסגרות ההצרכה החלקיות היכולות להופיע עימו. הקביעה נעשתה על פי בחינה באופן מדוקדק של כל פועל על ידי בוחן דובר עברית. למרות שקביעת מסגרת ההצרכה קשה לעיתים אף לדוברי השפה, ולא תמיד ברורה ההבחנה בין משלים מוצרך לתיאור, הרי שישנם משלימים שהקשר ביניהם לפועל הוא חזק דיו, כך שהדובר הילידי יכול לזהותם ביתר קלות.

קביעת המסגרות בצורה ידנית נעשתה מכיוון שאין בנמצא לקסיקון פעלים מקיף שניתן להשוות עימו. אמנם קיים הלקסיקון של שטרן (1994), אולם הוא מוגבל בהיקפו, ומרבית הפעלים המופיעים בהערכתנו אינם נמצאים בו כלל. כמו כן, לעיתים מסגרות ההצרכה המופיעות בו הן ארכאיות קמעה, ואינן משקפות את השפה הכתובה כיום. בנוסף, למרות שקיים בנק עצים בו יש ניתוח ידני של טקסטים, הניתוח אינו כולל, ברוב המקרים, הבחנה שלמה בין משלימים מוצרכים לתיאוריים, ועל כן שימוש בו להערכת איכות התוצאות עלול להניב תוצאות נמוכות.

בקביעת מסגרות ההצרכה לא נלקחו בחשבון צירופים פועליים כבולים אשר מסגרת ההצרכה ייחודית להם בלבד, כדוגמת ′מצא לנכון′ בו איננו מחשיבים את מילת היחס ′ל′ כחלק ממסגרת ההצרכה של הפועל **מצא**. בחלק הנספחים (נספח א) מופיעות, עבור כל פועל, מסגרות ההצרכה השונות המתאימות לו, על פי הניתוח הידני, ומובאת דוגמה מתוך טקסטים במשלב העיתונאי עבור כל מסגרת הצרכה שנקבעה.

6.1 הערכת התוצאות על ידי בחינת קבוצת פעלים בצורה ידנית

בחרנו להעריך את איכות תוצאות המערכת על ידי קבוצת פעלים המונה 17 פעלים. בחירת קבוצה זו של פעלים נעשתה תוך אותם שיקולים שהנחו את Briscoe ו-Carroll (1997) להערכת איכות עבודתם. פעלים אלו הם בעלי תפוצה מגוונת ושונים זה מזה במספר מסגרות ההצרכה שהם דורשים ובערכיות מסגרות ההצרכה. 17 הפעלים הם: **אהב**, **גרם**, **האמין**, **התחיל**, **התלונן**, **התעסק**, **חשב**, **כעס**, **מיהר**, **מצא**, **נראה**, **נתן**, **סיפק**, **עזר**, **ציפה**, **קיבל**, **שאל**.

ערכי הזהב עבור פעלים אלו התקבלו כאמור על ידי ניתוח ידני של הפעלים והם מוצגים בטבלה 6.1.

פועל	מסגרות הצרכה חלקיות
אהב	את, שם פועל, פסוקית
גרם	את, ל, שם פועל
האמין	ב, ל, על, פסוקית
התחיל	את, ב, עם, שם פועל
התלונן	ל, על, פסוקית
התעסק	ב, עם
חשב	על, שם פועל, פסוקית
כעס	על, ϕ
מיהר	אל, שם פועל

פועל	מסגרות הצרכה חלקיות
מצא	את, פסוקית
נראה	ב, ל, פסוקית
נתן	את, ל, שם פועל
סיפק	את, ל
עזר	ב, ל, שם פועל
ציפה	ל, מ, שם פועל, פסוקית
קיבל	את, מ, על
שאל	את, על

טבלה 6.1 : ערכי הזהב עבור קבוצת הפעלים הנבדקת

Table 6.1 : Gold standard for the test verbs

ביצועי כל אחד מהמדדים נבדקו על ידי השוואת מסגרות ההצרכה שבלקסיקון הפועל עבור כל אחד מ-17 הפעלים הנבדקים אל מול ערכי הזהב של מסגרות ההצרכה עבור פעלים אלו, והוערכו על ידי שני מדדים - דיוק (precision) ואחזור (recall). ערכי true positive (TP), true negative (TN), false positive (FP) ו-false negative (FN) מחושבים באופן הבא:

	זוג פועל-מסגרת הצרכה נמצא בלקסיקון	זוג פועל-מסגרת הצרכה לא נמצא בלקסיקון
זוג פועל-מסגרת הצרכה נמצא בין ערכי הזהב	true positive	false negative
זוג פועל-מסגרת הצרכה לא נמצא בין ערכי הזהב	false positive	true negative

טבלה 6.2 : הגדרת TP, TN, FP ו-FN

Table 6.2 : TP, TN, FP and FN definition

התוצאות שהתקבלו בהערכה פנימית זו מצביעות על כך שמדד ה-PMI הוא המדד המוצלח ביותר מבין ארבעת המדדים, בעל אחוז אחזור (recall) ודיוק (precision) הגבוהים ביותר (69.39% ו-97.14% בהתאמה). הערכים עבור שאר המדדים מופיעים בטבלה 6.3. אחוז הדיוק הגבוה מאוד מצביע על כך שרוב מסגרות ההצרכה המופיעות בלקסיקון הפועל של מדד ה-PMI הינן אכן מסגרות ההצרכה של הפועל, אולם אחוז האחזור מצביע על כך שקיימות מסגרות הצרכה של הפועל שהמדד לא הצליח לזהות.

מדד	TP	TN	FP	FN	Recall %	Precision %	F-score
PMI	34	120	1	15	69.39	97.14	80.95
***t*-score**	27	120	1	22	55.10	96.43	70.13
RF	29	103	18	20	59.18	61.70	60.42
LLR	17	100	19	34	33.33	47.22	39.08

טבלה 6.3 : תוצאות הערכת קבוצת הפעלים הנבדקת

Table 6.3 : Test verbs evaluation results

ניתן לראות את פירוט מסגרות ההצרכה שהתקבלו על ידי המערכת עבור הפעלים השונים, תחת המדדים השונים בטבלה 6.4 להלן.

פועל	LLR	PMI	RF	*t*-score
אהב	את	את, שם פועל	את, ϕ, שם פועל	את, שם פועל
גרם	ל, ϕ	ל	ל, ϕ	ל
האמין	את, ל, ϕ, פסוקית	ב, ל, פסוקית	ב, ל, ϕ, פסוקית	ב, ל, פסוקית
התחיל	ל, שם פועל, פסוקית	ב, עם, שם פועל	ב, ϕ, שם פועל	ב, שם פועל
התלונן	על	על, פסוקית	על, ϕ, פסוקית	על, פסוקית
התעסק	-	ב, עם	ב, עם, ϕ	ב, עם
חשב	את, ב, ל, ϕ, פסוקית	על, פסוקית	ϕ, פסוקית	פסוקית
כעס	-	על	על, ϕ	על
מיהר	שם פועל	אל, שם פועל	ϕ, שם פועל	שם פועל
מצא	את, שם פועל	את, פסוקית	את, ב, ϕ	את
נראה	-	ל, ϕ, פסוקית	ל, ϕ, פסוקית	ϕ, פסוקית
נתן	את, ב, ל, ϕ	את, ל	את, ל, ϕ	את, ל
סיפק	את	את, ל	את, ל, ϕ	את, ל

פועל	LLR	PMI	RF	*t*-score
עזר	ל	ל	ל,ϕ	ל
ציפה	מ, שם פועל	ל, מ, שם פועל, פסוקית	ל, מ,ϕ, פסוקית	ל, מ, פסוקית
קיבל	את, ב, ל, שם פועל, פסוקית	את, מ	את,ϕ	את
שאל	את, ב, ל, שם פועל	את	את,ϕ	את

טבלה 6.4 : תוצאות המערכת עבור קבוצת הפעלים הנבדקים

Table 6.4 : Test verbs results

6.2 דיון בתוצאות שהתקבלו

בסעיפים הבאים נקיים דיון על יכולות הזיהוי של מסגרות ההצרכה תחת המדדים הסטטיסטיים השונים, נסקור גורמים שונים שהשפיעו על התוצאות שהתקבלו ונבחן תופעות שונות שהופיעו בהן.

לאורך הסעיפים הבאים מופיעות טבלאות המייצגות את העיולים השונים בלקסיקון הפועל שהפקנו, עבור פעלים בקבוצת הבדיקה. לא כל הפעלים הופיעו בקורפורה עם כל מסגרות ההצרכה האפשריות, ועל כן חלק מהטבלאות כוללות שורות ריקות (עבור מקרים בהם לא היו נתונים). נתונים נוספים שאינם מפורטים מופיעים בהרחבה בחלק הנספחים (נספח ב).

נשתמש במוסכמות הבאות בטבלאות:

- מסגרות ההצרכה שאמורות היו להתקבל (על פי הניתוח הידני בחלק הנספחים) יודגשו על ידי ***מסגרת הצרכה*.**
- מקרי true positive - מסגרות שהתקבלו תחת מדד מסויים, ואכן היו אמורות להתקבל, יסומנו ב-(**TP**).
- מקרי false positive - מסגרות שהתקבלו תחת מדד מסויים, אך לא היו אמורות להתקבל, יסומנו ב-(**FP**).

6.2.1 מדד Raw Frequency

ההנחה היא, לכאורה, כי אם משלים הוא מוצרך אזי שכיחותו עם הפועל תהיה גבוהה בצורה משמעותית ביחס למשלימים האחרים, ולכן מדד זה יצליח לזהות את המשלימים המוצרכים. ואכן, בפעלים רבים התופעה מתקיימת, כדוגמת הפועל **תם**, אשר מסגרת ההצרכה היחידה שלו היא המסגרת הריקה. ההסתברות להופעתו עם מסגרת ההצרכה הריקה הינה כ-0.98, ועל כן מדד ה-raw frequency, כיתר המדדים, מצליח לזהותה בצורה מובהקת (ראה טבלה 9.1 בחלק הנספחים).

אולם, לעיתים ההסתברות הגבוהה משטה בנו. עבור הפועל **אמר** (טבלה 9.2 בחלק הנספחים) כ-55% מהמופעים הם מופעים עם מסגרת הצרכה ריקה. אולם, רוב המופעים הללו הם בשל נושא המופיע מיד לאחר הפועל, ולכן המדד מקבל פעמים רבות את מסגרת ההצרכה הריקה, למרות שאיננה מתאימה

לפועל. תופעה זו רווחת בקרב פעלים בהם הנושא נוטה להופיע מיד אחרי הפועל. כמו כן, עבור פעלים בעלי מספר מסגרות הצרכה שונות ייתכן כי תהיינה מסגרות הצרכה נכונות הקשורות לפועל אשר אחוז מופעיהן עם הפועל נמוך יחסית, ולכן לא יזוהו על ידי מדד זה. דוגמה לכך היא הפועל **החליט** (טבלה 9.3 בחלק הנספחים), בו מסגרות ההצרכה ״על״ ו״פסוקית״ לא מזוהות כלל בשל אחוז מופעים נמוך (6.64% ו-9.68% בהתאמה). בשל כך ישנה חשיבות לשימוש במדדים הנוספים, אשר מאפשרים לזהות את המסגרות הנכונות גם אם אינן נפוצות מספיק עם הפועל. בדוגמת הפועל **החליט** מדדי ה-t-score וה-PMI מצליחים לזהות את המסגרות הנכונות.

6.2.2 מדד Pointwise Mutual Information (PMI)

למדד ה-PMI יכולת טובה לזהות את ההצרכה בין פועל למשלימיו. דוגמה לכך ניתן לראות בפועל **נהר** (טבלה 9.4 בחלק הנספחים). רק 4% מהמופעים של הפועל **נהר** כוללים את מסגרת ההצרכה ״ל״ או ״אל״ מיד אחריו. זאת ממספר סיבות: ראשית, מופעים רבים שלו הם שגיאות בתיוג של שם העצם ״נהר״. כמו כן, מופעים רבים כוללים אחרי הפועל את הנושא, וכן תיאורי זמן, ולכן ההשלמה המוצרכת נמצאת הרחק מהפועל. בשל כך כ-91% ממופעי הפועל הינם מופעים עם מסגרת הצרכה ריקה. למרות זאת, מצליח מדד ה-PMI לזהות קשר של הצרכה בין הפועל למסגרת ההצרכה ״אל״, והוא היחיד המקבל מסגרת זו, ואף בניקוד גבוה יותר מזה של מסגרת ההצרכה הריקה (0.9952 לעומת 0.6876).

6.2.3 מדד *t*-score

כמו שטענו בחלק 3.3, מדד ה-*t*-score ומדד ה-PMI מזדהים מבחינת חיוביות הערכים, אולם מידת הזיקה בין הפועל למסגרת ההצרכה יכולה להיות שונה, ומכיוון שאנו נעזרים בחסם שאינו 0 על מנת לקבל הכרעה בדבר הקשר, קיים שוני בהכרעות בין המדדים השונים. בלקסיקון הפועל, כאשר מדד ה-t-score מכריע לחיוב, גם מדד ה-PMI מכריע כך, אך ההיפך אינו מתקיים. למרות זאת, מבחינת ההיררכיה בין המסגרות המתקבלות הם אינם תמיד מסכימים, דוגמה לכך היא הפועל **פונה** (טבלה 9.5 בחלק הנספחים). תחת מדד ה-PMI מסגרת ההצרכה ״מ״ מקבלת ניקוד גבוה יותר על פני מסגרת ההצרכה ״ל״ (1.1219 עבור ״מ״ לעומת 1.0833 עבור ״ל״), אך תחת מדד ה-*t*-score דווקא מסגרת ההצרכה ״ל״ מקבלת ניקוד גבוה יותר (0.4710 עבור ״ל״ לעומת 0.2293 עבור ״מ״). עם זאת, קיימים זוגות של פועל-מסגרת הצרכה נכונים, אותם מקבל מדד ה-PMI אך דוחה מדד ה-*t*-score. דוגמה לכך היא הפועל **נדבק** (טבלה 9.6 בחלק הנספחים). הפועל **נדבק** מופיע עם שבע מסגרות הצרכה שונות בקורפורה. מדדי ה-*t*-score וה-PMI מצליחים לזהות את מסגרות ההצרכה ״אל״, ״ב״ ו-״ל״ (עם הערכים הבאים בהתאמה: *t*-score: 0.8239, 0.5675, 0.23211, PMI: 2.9194, 1.1634, 0.6728) , אך מדד ה-PMI אף מצליח לזהות את מסגרת ההצרכה ״מ״ (עם הערך 0.4865).

6.2.4 מדד Log Likelihood Ratio (LLR)

התנהגות מדד ה-LLR אינה אחידה, וקשה להסיק ממנה על נכונות החלטות המדד. מצד אחד , למרות שהמדד מקבל רק 706 זוגות של פועל-מסגרת הצרכה, ונראה לכאורה כי הוא פוסל מסגרות הצרכה רבות, ונותר עם מסגרות הצרכה מועטות, אך נכונות, יש למדד ה-LLR נטייה לקבל יותר מדי מסגרות הצרכה, ללא יכולת להבחין בין המוצרכות ללא מוצרכות. דוגמה לכך ניתן לראות בפועל **נמצא** (טבלה 9.7 בחלק הנספחים). תחת מדד ה-LLR מתקבלות 5 מסגרות הצרכה שונות (את, ב, ל, שם פועל ופסוקית), מתוכן 3 מסגרות שגויות (את,ל, שם פועל), שאינן מתקבלות על ידי אף מדד אחר (מסגרות ההצרכה ′את′, ′ל′ ושם פועל), אולם גם מסגרת אחת נכונה שאינה מתקבלת על ידי אף מדד אחר (פסוקית). מאידך, ישנם פעלים בהם מתבצעת הפרדה ברורה מאוד בין המשלימים. זהו המקרה, למשל, בפועל **התגורר** (טבלה 9.8 בחלק הנספחים). הפועל **התגורר** מופיע בקורפורה 1,323 פעמים, עם 8 משלימים שונים. בפועל זה ניתן לראות כי מדד ה-LLR נותן ציון גבוה ביותר למסגרת ההצרכה ′ב′ לעומת שאר המשלימים (1,648.0065), וזוהי מסגרת ההצרכה היחידה שמתקבלת.

6.2.5 השפעת שכיחות הפועל בקורפורה על איכות התוצאות

שכיחות הפועל בקורפורה יכולה להשפיע על איכות התוצאות המתקבלות. מצד אחד, שכיחות גבוהה של פועל יכולה לגרום לרעש רב ולקבלת מסגרות הצרכה רבות שמהן יש לסנן את מסגרות ההצרכה הראויות. מנגד, עבור פעלים המופיעים בשכיחות נמוכה בקורפורה ייתכן כי הם אינם מופיעים עם כל מסגרות ההצרכה המתאימות להם. בסעיף זה נבחן ארבעה פעלים השונים זה מזה בשכיחותם בקורפורה. הפעלים שנבחן הינם **רצה** (שכיחות גבוהה), **נזקק** (שכיחות בינונית), **התנפל** ו-**התייעץ** (שכיחות נמוכה).

הפועל **רצה** מופיע כ-22,413 פעמים בקורפורה. הפועל מופיע עם כל המשלימים האפשריים, ולכן ישנה חשיבות רבה להבדלה בין המשלימים השונים. את יכולת ההפרדה של המדדים השונים ניתן לראות בטבלה 6.5. מסגרת ההצרכה שכל המדדים השונים מסכימים עליה, ונתנו לה את הניקוד הגבוה ביותר היא מסגרת שם הפועל. מסגרת נוספת המתקבלת רק על ידי המדדים t-score ו-PMI היא הפסוקית. ניתן לראות כי מדד ה-LLR מקבל משלימים רבים, אולם חלקם אינם מוצרכים כלל על ידי הפועל, כדוגמת מילת היחס ′ב′ או מילת היחס ′ל′, אך הוא היחיד המצליח לזהות את מסגרת ההצרכה ′את′. מדד ה-raw frequency מקבל בנוסף למסגרת ההצרכה הנכונה שם פועל, גם את המסגרת הריקה, שאינה מוצרכת על ידי הפועל. קבלת מסגרת הצרכה ריקה על ידי מדד זה נובעת ממופעי פועל אשר בהם הנושא מופיע מיד לאחר הפועל, ובשל כך לא מזוהה מסגרת ההצרכה הנכונה של הפועל, הנמצאת במקום אחר ביחס לפועל.

verb-frame	RF	LLR	*t*-score	PMI
רצה+אל	0.0001	164.9373	-0.0467	-4.0471
רצה+את	0.0565	(**TP**) 1828.0225	-0.1903	-0.9495
רצה+ב	0.0212	(**FP**) 2543.1029	-0.2098	-1.6259
רצה+ל	0.0108	(**FP**) 2937.0717	-0.2164	-2.1819
רצה+מ	0.0055	401.6401	-0.0839	-1.3802
רצה+על	0.0010	(**FP**) 1349.2493	-0.1363	-3.5458
רצה+עם	0.0006	144.2691	-0.0471	-2.1740
רצה	(**FP**) 0.1778	(**FP**) 7902.0989	-0.4229	-0.9462
רצה+שם פועל	(**TP**) 0.6622	(**TP**) 54416.5216	(**TP**) 1.8599	(**TP**) 2.2961
רצה+פסוקית	0.0645	3.1616	0.0089	0.0453

טבלה 6.5 : הפועל 'רצה'

Table 6.5 : '*rch*'

הפועל **נזקק** (טבלה 6.6) מופיע בקורפורה כ-549 פעמים, עם 8 משלימים שונים. כל המדדים מצליחים לזהות את מסגרת ההצרכה היחידה 'ל', אולם מדד ה-raw frequency מקבל גם את המסגרת הריקה. הסיבה לכך יכולה להיות נעוצה בעובדה כי נטיות שונות של הפועל, כדוגמת נזקק או נזקקים, הינם שמות עצם, והסיבה למנייתם היא בשל טעויות של המתייג המורפולוגי.

verb-frame	RF	LLR	*t*-score	PMI
נזקק+אל	-	-	-	-
נזקק+את	0.0142	118.8942	-0.2766	-2.3258
נזקק+ב	0.0163	71.6506	-0.2188	-1.8852
נזקק+ל	(**TP**) 0.6414	(**TP**) 978.9799	(**TP**) 1.3792	(**TP**) 1.9054
נזקק+מ	0.0018	17.2101	-0.1014	-2.4828
נזקק+על	0.0036	24.7152	-0.1237	-2.2275
נזקק+עם	-	-	-	-
נזקק	(**FP**) 0.2727	79.5624	-0.2756	-0.5181
נזקק+שם פועל	0.0442	4.9735	-0.0665	-0.4087
נזקק+פסוקית	0.0054	49.0151	-0.1732	-2.4238

טבלה 6.6 : הפועל 'נזקק'

Table 6.6 : '*nzqq*'

הפועל **התנפל** (טבלה 6.7) מופיע 164 פעמים בקורפורה, עם 4 משלימים שונים: 'את', 'ב', 'על', והמסגרת הריקה. למרות שהפועל מופיע רק מספר פעמים מועט בקורפורה, ניתן לראות כי כל המדדים הצליחו לזהות בצורה ברורה את מסגרת ההצרכה הנכונה 'על'. מדד ה-raw frequency מקבל בנוסף גם את מסגרת ההצרכה הריקה, מאותן סיבות שמנינו לעיל.

verb-frame	RF	LLR	*t*-score	PMI
התנפל+אל	-	-	-	-
התנפל+את	0.0061	43.0835	-0.2973	-3.1753
התנפל+ב	0.0183	20.1775	-0.2142	-1.7748
התנפל+ל	-	-	-	-
התנפל+מ	-	-	-	-
התנפל+על	**(TP)** 0.7561	**(TP)** 661.0187	**(TP)** 2.9674	**(TP)** 3.1085
התנפל+עם	-	-	-	-
התנפל	**(FP)** 0.2195	40.3657	-0.3548	-0.7352
התנפל+שם פועל	-	-	-	-
התנפל+פסוקית	-	-	-	-

טבלה 6.7 : הפועל 'התנפל'

Table 6.7 : '*htnpl*'

מנגד, ניתן לראות כי בשל מספר מופעים קטן של הפועל וההגבלה על מיקום המשלים אנו יכולים, לעיתים, שלא לקבל את התמונה המלאה של הפועל. דוגמה לכך היא הפועל **התייעץ** (טבלה 6.8). הפועל **התייעץ** מופיע 217 פעמים בקורפורה, עם 5 משלימים שונים: 'את', 'ב', 'על', 'עם' והמסגרת הריקה. ניתן לראות כי כל המדדים הצליחו לזהות בצורה מובהקת את מסגרת ההצרכה 'עם'. למרות זאת, מסגרת ההצרכה הנוספת היכולה להופיע עם הפועל היא 'על', אולם מכיוון שרוב המופעים של מסגרת זו הם בקורפורה הם מופעים של מסגרת הצרכה רחבה יותר, המאופיינת בסדר יחסית נוקשה - התייעץ + עם + על, מופעי המסגרת אינם באים לידי ביטוי בלקסיקון. כמו כן שוב בולטת התופעה של מסגרת ההצרכה הריקה המתקבלת רק על ידי מדד ה-raw frequency.

verb-frame	RF	LLR	*t*-score	PMI
התייעץ+אל	-	-	-	-
התייעץ+את	0.0876	6.7703	-0.1226	-0.5109
התייעץ+ב	0.0230	23.0954	-0.2029	-1.5440
התייעץ+ל	-	-	-	-
התייעץ+מ	-	-	-	-
התייעץ+על	0.0092	5.5975	-0.1008	-1.2986
התייעץ+עם	**(TP)** 0.4378	**(TP)** 712.0832	**(TP)** 4.5686	**(TP)** 4.4758
התייעץ	**(FP)** 0.4424	0.2102	-0.0231	-0.0344
התייעץ+שם פועל	-	-	-	-
התייעץ+פסוקית	-	-	-	-

טבלה 6.8 : הפועל 'התייעץ'

Table 6.8 : '*htii'c*'

6.2.6 השפעת צירופים פועליים כבולים על היכולת לזהות את מסגרות ההצרכה

גורמים שונים הנעוצים בהתנהגות הפועל משפיעים על יכולתנו לזהות נכונה את מסגרת ההצרכה של הפועל. כבר בדיון על מדד ה-raw frequency (סעיף 5.2.1) ציינו כי ישנם פעלים, כדוגמת הפועל **אמר**, הנוטים לסדר VSO במשלב העיתונאי אליו שייכים הקורפורה. גורם נוסף היכול להשפיע על הזיהוי הנכון הוא מופעי צירופים פועליים כבולים. צירופים פועליים כבולים הינם צירופים כמו *׳התחנן על נפשו׳*, בהם יש קשר חזק הן בין הפועל למילת היחס, והן בין מילת היחס לשם העצם. צירופים כאלו יכולים לגלם ייחודיות במשמעותם עם שם עצם מסוים, אולם הם יכולים גם לגרום להגבלה בקבוצת שמות העצם היכולה להופיע עם מילת היחס. כזהו הפועל **הועמד** (טבלה 9.9 בחלק הנספחים). הפועל **הועמד** יכול לקבל את מסגרת ההצרכה ׳ל׳ רק במקרים ספורים, כדוגמת בצירוף צירופים שמניים מתחום המשפטי כדוגמת ׳משפט׳ או ׳דין׳ או בצירוף נטיות ׳רשות׳ (*׳הועמד לרשותו׳*), מכיוון שמדובר בצירוף פועלי כבול. למרות זאת, מכיוון שעבודתנו מתרכזת רק בקשר בין הפועל למשלים עצמו, ולא בקשר בין המשלים לשם העצם הבא אחריו, המערכת לא מבדילה בייחודיות זו, ומכלילה את מופעי הפועל **הועמד** עם המשלים ׳ל׳ לכדי מסגרת הצרכה, ללא האלמנט הנוסף הנדרש במסגרת זו. כל המדדים זיהו את ׳ל׳ כמסגרת הצרכה, ובנוסף הצליח מדד ה-PMI לזהות את מסגרת ההצרכה ׳על׳.

6.2.7 זיהוי מסגרות הצרכה עבור מסגרות רחבות בנות יותר ממשלים אחד

כמו שציינו בפרק 1, לפועל יכולות להיות מסגרות הצרכה בנות יותר מארגומנט אחד. בשל סדר המשלימים החופשי יחסית בעברית, ציפינו כי במקרה של מסגרת הצרכה רחבה יתקבלו המשלימים השונים כמסגרות הצרכה חלקיות. ואכן ניתן לראות כי הדבר מתקיים במקרים רבים, כמו בפועל **הביא** (טבלה 9.10 בחלק הנספחים), לו מסגרת הצרכה הכוללת שני ארגומנטים: ׳ל׳ ו-׳את׳. כל המדדים מצליחים לזהות את שני המשלימים הללו כמוצרכים. לעומתו, ניתן להתבונן בפועל **איחל** (טבלה 9.11 בחלק הנספחים). מסגרת ההצרכה היחידה שזוהתה באופן מובהק על ידי כל המדדים היא ׳ל׳, אך אף אחד מהמדדים לא זיהה את מסגרת ההצרכה של הצירוף השמני, זאת למרות שמסגרת ההצרכה של הפועל כוללת את מילת היחס ׳ל׳, וצירוף שמני, כדוגמת:

(א) אולמרט **איחל** לגרנט הצלחה. (הארץ, 23.10.2007)

ניתן להסביר זאת בכך שבקורפורה הנבדקים הפועל **איחל** מופיע עם מסגרת שאינה גמישה: איחל + ל + צירוף שמני. הסיבה לכך יכולה להיות נעוצה בתכונת הפועל, או בתכונת הקורפורה עצמם, או שמא בשל העובדה שרוב המופעים כללו את מילת היחס ׳ל׳ בנטיית כינוי גוף ולכן ארגומנט זה הופיע ראשון. תהא הסיבה אשר תהא, הדבר מדגיש את העובדה כי מסגרות ההצרכה שנלמדו הינן חלקיות בלבד.

6.2.8 השפעת איכות הניתוח המורפולוגי על תוצאות תהליך זיהוי מסגרות ההצרכה

אנו משתמשים בקורפורה המנותחים מורפולוגית, דבר הנדרש לצורך זיהוי נכון של הפועל ומסגרות ההצרכה שלו, בעיקר בשפה העברית, בשל מורכבותה המורפולוגית. אך תוצאות המנתח המורפולוגי יכולות להשפיע רבות על תוצאותינו. מכיוון שאנו עוסקים בטקסט שאינו מנוקד קיימים פעלים רבים עבורם לא ניתן לקבוע מהו הניתוח הנכון. כזו היא צורת העתיד של הפעלים **אילץ** ו-**נאלץ**. שני הפעלים מזדהים בצורת העתיד שלהם, ומכיוון שכפי שציינו בסעיף 3.1.1, המנתח המורפולוגי אינו מכריע בין ניתוחי הפועל האפשריים, אנו מקבלים את שני הניתוחים. יתרה מזו, מכיוון שאנו משתמשים בטכניקת עידון הציונים (סעיף 3.1.1), התלויה בין היתר בתפוצת הבניין, יש נטייה להעדפת הניתוח של בניין פיעל על פני נפעל, ובשל כך מופעי הפועל **נאלץ** משפיעים מאוד על מסגרות ההצרכה של הפועל **אילץ**.

הפועל **אילץ** (טבלה 9.12 בחלק הנספחים) מופיע בקורפורה עם 8 מסגרות הצרכה שונות. מסגרת ההצרכה של הפועל בנויה משני ארגומנטים: ′את′ ושם פועל, ואכן, ניתן לראות כי כל המדדים קיבלו את שני הארגומנטים הללו. אולם בחינה של מופעי הפועל **אילץ** מעלה כי לא קיימים מופעים אשר בהם מיד לאחר הפועל מופיע שם פועל. הסיבה לכך היא כי במסגרות ההצרכה הכוללות שם פועל ומשלים נוסף, נוטה שם הפועל להופיע אחרון, והגמישות בין הארגומנטים במסגרת ההצרכה נחלשת. זוהי תכונה של השפה המנסה למזער את בעיית עמימות התאמת המשלים בין הפועל לשם הפועל. לעומת זאת, מסגרת ההצרכה של הפועל **נאלץ** (טבלה 9.13 בחלק הנספחים) כוללת רק את שם הפועל ובשל העדפת הניתוח לבניין פיעל קיבלנו את מסגרת ההצרכה שם פועל לפועל **אילץ**.

לעומת זאת, עבור הפועל **ניצֵל** (טבלה 9.14 בחלק הנספחים) דווקא העדפת בניין פיעל על פני נפעל היא זו המונעת את קבלת מסגרת ההצרכה ′מ′ (הנפוצה עם הפועל ניצַל (טבלה 9.15 בחלק הנספחים)).

6.2.9 השפעת הקורפוס הנבדק על תוצאות תהליך זיהוי מסגרות ההצרכה

ישנם פעלים רבים שהתנהגותם תלויה במשלב בו הם מופיעים. דוגמה לכך הוא הפועל **מחה** (טבלה 9.16 בחלק הנספחים). הפועל **מחה** יכול לקבל הן את מסגרת ההצרכה ′על′ והן את מסגרת ההצרכה ′את′ (כדוגמת ′מחה אותם מעל פני האדמה′), אולם, המופעים עם מסגרת ההצרכה של צירוף שמני היא נדירים מאוד בטקסטים, ועל כן אף אחד מהמדדים אינו מצליח לזהותם. מנגד, את מסגרת ההצרכה ′על′ כל המדדים מצליחים לזהות בבירור.

6.3 בעיית הצמדת צירוף יחס (pp-attachment)

6.3.1 תיאור הבעיה

צירוף היחס הוא רכיב תחבירי בעל גמישות רבה בהצטרפותו. הוא יכול להצטרף הן לפועל והן לשם עצם, כתלות בגורמים שונים, תחביריים, סמנטיים ופרגמטיים. בעיית ההצמדה נפוצה מאוד בניתוח

תחבירי, והיא בעלת השפעה רבה בניתוח אוטומטי של טקסט, ובשל כך יש לה משקל מכריע בגרימת טעויות בניתוח (Lin, 1998).

למרות שתיאורטית אפשרויות ההצמדה של צירוף היחס הן רבות, ותלויות במספר שמות העצם והפעלים המופיעים במשפט, רוב המחקר שנעשה בתחום הוא על מבנה תחבירי נפוץ, בו צירוף היחס מופיע מיד לאחר צירוף שמני, כאשר הצירוף השמני מופיע מיד לאחר פועל (Hindle & Rooth, 1993).

שיטות שונות פותחו במהלך שני העשורים האחרונים בניסיון לזהות את ההצטרפות הנכונה, ביניהן שיטות סטטיסטיות טהורות המשתמשות בשכיחויות מופעי הפועל, שם העצם ומילת היחס (Ratnaparkhi et al., 1994; Hindle & Rooth, 1993; Resnik & Hearst, 1993) וכאלו המשתמשות בנוסף במידע סמנטי (Hirst, 1988; Jensen & Binot, 1987; Dahlgren & McDowell, 1986; Wilks et al., 1985) .

בשל התלות החזקה של איכות הניתוח התחבירי ב-pp-attachment נעשו אף ניסיונות לפתח מנתחים שישלבו את המידע על הקשר בין מילות היחס לפועל בתהליך הניתוח התחבירי (Atterer & Schütze, 2007; Foth & Menzel, 2006).

אחת ההתניות התחביריות היכולה להיות לעזר בזיהוי ההצטרפות הנכונה היא מסגרת ההצרכה של הפועל. כאשר פועל דורש השלמה מסוימת, וזו יכולה להתגלם אך ורק בצירוף היחס בעל העמימות התחבירית, סביר יותר שצירוף היחס קשור לפועל, שכן אחרת מופע הפועל במשפט אינו שלם. ניתן לראות זאת בדוגמה הבאה:

(א) לא נמצאו רשיונות לנשיאת נשק

(ב) הוא הראה רשיונות לשוטר

בשני המשפטים מופיעה תבנית זהה - פועל, שם העצם **רשיונות**, מילת היחס **ל**, כאשר ההבדל בין המשפטים הוא הפועל המשובץ בכל תבנית. במשפט (א) הפועל **נמצא** אינו כולל את מילת היחס **ל** בין מסגרות ההצרכה שלו, ואילו במשפט (ב) הפועל **הראה** דורש את מילת היחס **ל**, ועל כן במשפט (א) מוצמדת מילת היחס לשם העצם, ואילו במשפט (ב) מוצמדת מילת היחס לפועל.

זוהי התניה המגיעה רק מכיוון הפועל הנידון, ולשם העצם אין השפעה על ההחלטה, אולם ניתן לראות כי הוספת המידע על ההתניה התחבירית של הפועל, למרות שאינה כוללת את כלל השיקולים בבחירת ההצטרפות, משפרת רבות את איכות ההצטרפות (Volk, 2002; Pantel & Lin, 2000; Yeh & Vilain, 1998; Stetina & Nagao, 1997). בשל כך אנו משתמשים בלקסיקון שקיבלנו להוספת מימד בבחירת הצטרפות צירוף היחס, מתוך הנחה כי השגת שיפור באופן הזה בהצטרפות מאששת את נכונות הלקסיקון.

למרות זאת, קיימים גורמים נוספים המשפיעים על הבחירה הנכונה. ייתכן כי במשפט הנבדק קיימת ההשלמה המוצרכת לפועל דווקא בסביבה רחבה יותר של הפועל, אותה אין אנו בוחנים. כמו כן, בשל האפשרות של מספר מסגרות הצרכה לפועל, ייתכן כי קיימת מסגרת הצרכה של הפועל שמילת היחס המופיעה היא חלק ממנה, אך דווקא במופע הנבדק המימוש של הפועל הוא שונה, ובעזרת מסגרת הצרכה אחרת, ועל כן מילת היחס לא מצטרפת לפועל. למרות הגורמים הללו היכולים להפוך את ההחלטה שבוצעה על סמך מסגרת ההצרכה בלבד ללא נכונה, מצאנו כי הוספת המידע הובילה לשיפור באיכות ההצמדה, והדבר מאשש את איכות התוצאות.

6.3.2 שיטת ההערכה

על מנת להעריך את השיטה ואת המדדים השונים בהם השתמשנו לצורך למידת מסגרת ההצרכה של הפועל, אנו משתמשים בנתונים שקיבלנו כדי להכריע בבעיית הצמדת צירופי יחס (6.3.1). המבנה בו בחרנו להתמקד הוא מבנה בו צירוף היחס מופיע מיד לאחר צירוף שמני, שמופיע מיד לאחר פועל (6.3.1). באופן מדויק יותר, המבנה אותו בדקנו הוא מבנה מהצורה v - n - p כאשר בין הפועל לשם העצם, ובין שם העצם למילת היחס אנו מתירים הופעות של מילים שונות שאינן פועל, שם עצם או מילת יחס. הרציונל מאחורי בחירת מבנה זה הוא כי בו מגולמת בוודאות העמימות, שכן מילת היחס יכולה להיות בשני תפקידים – או בתפקיד מושאי ביחס לפועל, *v*, או בתפקיד לוואי של השם, *n*.

חלק המבחן בו השתמשנו מהווה 80% מהמשפטים בבנק העצים, וכולל 4,224 משפטים, בהם 1,301 פעלים ו-6,025 מופעי פועל, ובו מצאנו 323 מופעים מהמבנה הנ"ל, עם 204 פעלים שונים.

הערכת כל אחד מארבעת הלקסיקונים (תוצאות ארבעת המדדים הסטטיסטיים שבחנו) נעשתה באופן הבא: עבור כל שלשה (v,n,p) ההכרעה האם מילת היחס *p* קשורה לפועל, *v*, או לשם העצם, *n*, נעשתה על סמך המידע שבלקסיקון: אם מילת היחס הופיעה כמסגרת הצרכה של הפועל בלקסיקון, אזי היא הוצמדה לפועל, אחרת היא הוצמדה לשם. את הקביעות הללו השווינו להצמדה הנכונה הידועה מבנק העצים. נציין כי השימוש במידע הנמצא בלקסיקון נעשה רק עבור פעלים המופיעים יותר מ-100 פעמים בקורפורה. עבור פעלים עם מספר מופעים קטן יותר ההכרעה בדבר הקשר בין מילת היחס לפועל היא תמיד שלילית.

תוצאות הניסויים עבור ארבעת הלקסיקונים הושוו לבסיס (baseline) בו קבענו כי צירוף היחס לעולם לא יצטרף לפועל. בנוסף, ניסינו לשלב בין המדדים השונים, ובחנו את התוצאות עבור Voting, בו ההכרעה על הצמדת צירוף היחס לפועל נעשית על פי הכרעת הרוב בארבעת הלקסיקונים של המדדים הסטטיסטיים השונים, וכן OR, בו ההכרעה על ההצמדה לפועל דורשת הסכמה של לקסיקון אחד לפחות. בטבלה 6.10 ניתן לראות את תוצאות הניסויים שערכנו עם ארבעת הלקסיקונים LLR, PMI, raw frequency ו-*t*-score, וכן עבור הניסויים Voting ו-OR. בטבלה מופיעים ערכי

ה-(TP) true positive, (TN) true negative, (FP) false positive ו-(FN) false negative, עבור כל מדד המחושבים באופן המתואר בטבלה 6.9, וכן מדד הדיוק (accuracy), וההפחתה בשיעור הטעות (error reduction ratio, ERR).

	מילת היחס מצטרפת לפועל על פי הלקסיקון	מילת היחס לא מצטרפת לפועל על פי הלקסיקון
מילת היחס מצטרפת לפועל על פי בנק העצים	true positive	false negative
מילת היחס לא מצטרפת לפועל על פי בנק העצים	false positive	true negative

טבלה 6.9 : הגדרת TP, TN, FP ו-FN עבור בעיית הצמדת צירוף יחס

Table 6.9 : TP, TN, FP and FN definition for the PP-attachment problem

הדיוק (accuracy) מחושב באופן הבא:

$$\text{accuracy} = \frac{TP+TN}{TP+FP+TN+FP} \tag{6.1}$$

בחינת הנתונים (טבלה 6.10) מראה כי תחת שיטת הערכה זו מדד ה-*t*-score משיג את התוצאות הטובות ביותר (הפחתה של 28.85% בשיעור הטעויות לעומת ה-baseline) לעומת שאר המדדים הנבחנים.

Method	FN	FP	TN	TP	Accuracy	ERR
Baseline_FALSE	156	0	167	0	51.70%	
LLR	108	31	136	48	56.97%	10.90%
PMI	72	52	115	84	61.61%	20.51%
RF	86	35	132	70	62.54%	22.44%
***t*-score**	**84**	**27**	**140**	**72**	**65.63%**	**28.85%**
Voting	92	26	141	64	63.47%	24.36%
OR	61	69	98	95	59.75%	16.67%

טבלה 6.10 : תוצאות הצמדת מילות היחס

Table 6.10 : PP-attachment results

באופן זהה ניסינו לבחון את השיטה עבור שלשות של פועל-שם עצם-׳ש׳ או ׳כי׳. ׳ש׳ ו-׳כי׳ יכולים להצטרף הן לפועל (כמוצג בפרק 1) והן לשם העצם כמו בדוגמאות הבאות:

(א) ביסודו של המנגנון **מונחת** התפיסה כי על עורך הדין להיות אמון על שמירת החוק. (הארץ, 24.4.06)

(ב) בבזק **שורר** חשש כי כישלון המכרז יהווה איתות שלילי לשוק ההון. (הארץ, 25.7.03)

(ג) אף פעם לא **נשמעה** טענה ש'מה פתאום נשיא בית המשפט העליון מנהל את בתי המשפט'. (ידיעות אחרונות, 12.11.07)

את השלשות הללו הגבלנו לכאלה בהם שם העצם מופיע מיד לאחר הפועל, והמיליות 'ש' או 'כי' מופיעות מיד לאחר שם העצם, על מנת למנוע מגורמים נוספים היכולים להשפיע על ההצטרפות להופיע בתווך.

בחלק המבחן מבנק העצים נמצאו 20 מופעים במבנה הנ"ל, עם 19 פעלים שונים. בטבלה 6.11 ניתן לראות את תוצאות הניסויים שערכנו עם ארבעת הלקסיקונים LLR, PMI, raw frequency ו-*t*-score, וכן עבור הניסויים Voting ו-OR. בחינת הנתונים מראה כי תחת הערכה זו מדד ה-raw frequency משיג את התוצאות הטובות ביותר (הפחתה של 75% בשיעור הטעויות לעומת ה-baseline).

נשים לב כי המדדים השונים הצליחו לזהות את כל המקרים בהם ההצטרפות הייתה לפועל. כמו כן, אותם מקרים בהם מדד ה-PMI ומדד ה-*t*-score הכריעו עבור הצטרפות לפועל, אך למעשה ההצטרפות הנכונה הייתה לשם העצם, הם מקרים בהם הופיעו פעלים עבורם האפשרות של הצטרפות פסוקית קיימת, אולם לא בשילוב עם ההשלמה השמנית, ומכיוון שהידע הזה לא קיים בלקסיקון לא היתה סיבה להכריע אחרת. מקרים אלו הם השלשות הבאות (בסוגריים מופיעים המדדים שקיבלו אותן):

(א) ראינו מועצות ש - ראה ש (PMI, *t*-score)

(ב) קבענו עיקרון ש - קבע ש (LLR, PMI, RF, *t*-score)

זוהי הסיבה לכך שדווקא מדד ה-RF הניב ERR גבוה יותר על פני PMI ו-*t*-score. ההבדל ביניהם הוא רק בהכרעה על השלשה 'ראינו מועצות ש', אשר בה מכריע RF כי 'ש' אינו יכול להצטרף לפועל **ראה**. למרות שההכרעה נכונה לשלשה המסוימת הזו, הקביעה אינה נכונה עבור הפועל, שכן הפועל **ראה** יכול לקבל כמשלים פסוקית.

Method	FN	FP	TN	TP	Accuracy	ERR
Baseline_FALSE	4	0	16	0	80.00%	
LLR	0	4	12	4	80.00%	0.00%
PMI	0	2	14	4	90.00%	50.00%
RF	**0**	**1**	**15**	**4**	**95.00%**	**75.00%**
t-score	0	2	14	4	90.00%	50.00%
Voting	**0**	**1**	**15**	**4**	**95.00%**	**75.00%**
OR	0	5	11	4	75.00%	-25.00%

טבלה 6.11 : תוצאות הצמדת הפסוקיות

Table 6.11 : Complementizer-attachment results

מיצוע התוצאות שהתקבלו על פני שני סוגי ההצמדות השונות מראה כי תחת שיטת הערכה זו דווקא מדד ה-raw frequency (מתוך ארבעת המדדים הנבדקים) מספק את התוצאות הטובות ביותר, עם דיוק (accuracy) ממוצע של 78.77% (טבלה 6.12), אך שימוש בהכרעת הרוב משפר במעט את התוצאות, עם דיוק (accuracy) של 79.38%.

Method	Accuracy
LLR	68.49%
PMI	75.81%
RF	**78.77%**
t-score	77.82%
Voting	79.24%
OR	67.38%

טבלה 6.12 : ממוצע accuracy של המדדים השונים

Table 6.12 : Accuracy average for the statistical tests

6.4 תרגום אוטומטי

6.4.1 תיאור הבעיה

בתרגום אוטומטי נתקלים לעיתים בבעייתיות בתרגום מילות היחס, בעיקר אלו המהוות חלק ממסגרת ההצרכה של הפועל. מילות יחס כאלה, לעומת מילות יחס תיאוריות, הן בדרך כלל חסרות משמעות, והופעתן נקבעת שרירותית על ידי הפועל הבוחר אותן. במעבר בין שפות שונות ניתן לראות כי קיימים פעלים בשפה האחת בעלי מילת יחס מוצרכת כאשר בשפה האחרת לא מופיעה כלל מילת יחס, כדוגמת הפועל **להשתמש ב** לעומת **use**, באנגלית, ופעלים בעלי מילת יחס מוצרכת שהמקבילה להם הוא אמנם פועל עם מילת יחס, אך מילת היחס איננה התרגום הנפוץ, כדוגמת **לדאוג ל** לעומת **worry about**. בשל כך מיפוי של מילת יחס אחת לשנייה לא יועיל, ויש צורך בטיפול מיוחד במילות היחס. אחת הדרכים היא להשתמש בידע הלקסיקלי שבידנו, על מסגרות ההצרכה של הפועל על מנת ליצור ולבחור את התרגום הנכון.

6.4.2 שיטת ההערכה

אימות הלקסיקון שהפקנו נעשה על ידי שילובו במערכת לתרגום ערבית-עברית[2] (Shilon et al., 2012a).

[2] תודתנו לרשף שילון, ניזאר חבש, אלון לביא ושולי וינטנר על האפשרות להשתמש במערכת התרגום לצורך הערכת מחקרנו.

למרות שעברית וערבית הן שפות בעלות מקור משותף, וחלק ממילות היחס דומות, התופעות שהצגנו לעיל לא נעדרות מן השפות הללו, ואף קיימות מילות יחס בשפה האחת שאין להם מקבילה בשפה האחרת, כדוגמת מילת היחס **את** בעברית.

על מנת להשתמש בנתוני הלקסיקון שהופקו נקבע באופן ידני סף קבלה על הניקוד שקיבלו זוגות הפועל-משלים, ונלקחו רק 2,325 זוגות, ביניהם 1,402 פעלים שונים. כמו כן בעבודתם של שילון ואחרים (2012a) נעשה שימוש רק בנתוני הלקסיקון הנוגעים למילות היחס ב, ל, מ, על ועם, ולמשלים השמני.

הגישה בה השתמש שילון בלקסיקון שהפקנו היא ראשית לתרגם את מילות היחס בשפת המקור (ערבית) לקבוצה מצומצמת של מילות יחס, ולאחר מכן להוסיף את ההגבלות של מילות היחס על הפעלים. המידע מהלקסיקון שולב בשני חלקים של המערכת. החלק הראשון הוא המחולל המורפולוגי לעברית, בו המידע שולב כחלק מתכונות כל פועל. בנוסף, המידע הוכנס לחלק הדקדוק, כהוספת אילוצים בין הפועל לבין משלימיו האפשריים. שילון מאפשר ארבעה מיפויים תחביריים: מיפוי של צירוף שמני לצירוף שמני (אין הוספת מיליות, למעט ׳את׳ במקרה המיודע), מיפוי של צירוף שמני לצירוף יחס (הוספת מילת יחס מתאימה בצד העברית), מיפוי של צירוף יחס לצירוף שמני (השמטת מילת היחס שהופיעה בערבית), ומיפוי של צירוף יחס לצירוף יחס (מילת היחס אינה בהכרח מיפוי ישיר של מילת היחס שהופיעה בצד הערבית). במיפוי של צירוף שמני לצירוף יחס המיפוי נעשה לכל מילות היחס, שכן אין לדעת אילו מבין מילות היחס תתאים, ועבור המיפוי של צירוף יחס לצירוף יחס, המיפוי נעשה למילות יחס אפשריות, על פי מילון דו-לשוני קיים. בשלב הדקדוק מוצפות מילות היחס האפשריות, והבחירה ביניהן נעשית ברמת המשפט, על פי החיתוך עם האילוצים של כל פועל.

השיטה נבדקה על 28 משפטים קצרים (בני לכל היותר 10 מילים), אל מול בסיס (baseline) בו הושמטו האילוצים על הפעלים, כך שכל פועל יכול לקבל כל מילת יחס, וכן הלקסיקון הדו-לשוני הכיל רק את המיפויים הנפוצים בין מילות היחס בשפות.

התוצאות המופיעות בטבלה 6.13 מראות שיפור בעת שימוש בנתוני הלקסיקון שהפקנו. לפרטים נוספים בדבר מערכת התרגום ושילוב מסגרות ההצרכה ראו Shilon et al. (2012b).

	BLEU	METEOR
המערכת עם האילוצים על מילות היחס	37.0	56.0
המערכת ללא אילוצים על מילות היחס	32.5	52.6

טבלה 6.13 : הערכת איכות התרגום

Table 6.13 : Translation evaluation

פרק 7

למידת מסגרות הצרכה מתוך קורפוס מתויג מורפולוגית ומנותח תחבירית

בשלב הראשון של המחקר, למידת מסגרות ההצרכה נעשתה על פי לקיחת המשלים הראשון המגיע מיד לאחר הפועל, מתוך הנחה כי בשפה קיימת נטייה להקדים את העיקר לעומת הטפל, ולכן משלים יופיע לפני תיאורים שונים (פרק 5). גישה זו התפתחה בשל הצורך למצוא את המשלימים הפוטנציאלים לפועל, בלא לדעת את מיקומם, זאת מכיוון שבתחילת מחקרנו לא היה בנמצא מנתח תחבירי לעברית.

תפיסה זו של חיפוש מסגרת ההצרכה מיד לאחר הפועל היא מעט בעייתית. קיימים פעלים עבורם מסגרת ההצרכה היא רחבה יותר, כדוגמת הפועל **איחל** שאחת ממסגרות ההצרכה שלו היא בת שני ארגומנטים וכוללת את מילת היחס ״ל״ וצירוף שמני. כמו שציינו בסעיף 6.2.7, למרות שבשל תכונת העברית, הסדר בין המשלימים אינו קבוע בהכרח, עדיין ייתכן כי תחת תנאים מסויימים (כגון משלב או סגנון) הופעת מסגרת ההצרכה של פועל תהיה בעלת סדר קבוע. בשל כך לקיחת המשלים הראשון יכולה לגרום להתעלמות מיתר המשלימים, ולאי קבלת התמונה השלמה לגבי מסגרת ההצרכה. בדוגמת הפועל **איחל**, אכן התקבלה מסגרת הצרכה חלקית בלבד, הכוללת את מילת היחס ״ל״, אך הצירוף השמני לא התקבל כמשלים אפשרי.

לכן, השלב הבא בו בחרנו להמשיך את מחקרנו הוא הרחבת הנתונים שלנו ממשלים הקרוב ביותר לפועל למשלים הקשור לפועל, בעזרת המנתח התחבירי שנעשה במסגרת עבודת הדוקטורט של יואב גולדברג (Goldberg, 2011). השימוש במנתח תחבירי לצורך למידת מסגרת ההצרכה הוא טבעי ומתבקש, אך בתחילת מחקרנו לא היה כלל מנתח תחבירי לעברית, ולכן פיתחנו שיטות אלטרנטיביות במטרה להתגבר על משוכה זו. כמו כן ישנה חשיבות במחקרנו להשוואה בין השיטות השונות, על מנת לבחון האם אכן הוספת ידע תחבירי מועילה במידה מרובה לתהליך הלמידה, או שמא דווקא הטכניקה הפשטנית לכאורה, המסתכלת רק על סביבה קרובה מאוד לפועל, יכולה להספיק.

יש לזכור כי כמו כל כלי חישובי, גם המנתח התחבירי יכול לטעות בניתוחו, ולכך השפעה רבה על תהליך למידת מסגרות ההצרכה. על פי גולדברג, רמת הדיוק של המנתח התחבירי היא 79.5%. כמו כן נדגיש כי המנתח התחבירי משתמש במתייג מורפולוגי שפותח על ידי מני אדלר (Adler, 2007), שהוא שונה מהמתייג המורפולוגי בו השתמשנו בחלקה הראשון של עבודתנו (המתייג המורפולוגי של מיל"ה), ואף שינוי זה יכול להשפיע על איכות התוצאות.

המנתח התחבירי נותן עבור כל פועל קבוצת מילים הקשורה אליו. אנו בחרנו לא להשתמש בקבוצה זו כיחידה שלמה, אלא לראות כל משלים המופיע עם הפועל כיחידה נפרדת. טכניקה זו משמרת את הגישה של חלקו הראשון של מחקרנו בדבר למידת מסגרות ההצרכה מתוך קורפוס מתויג מורפולוגית ואף היא מאפשרת לנו ללמוד את מסגרות ההצרכה החלקיות בלבד של הפועל.

הקורפוס בו השתמשנו בשיטה זו הינו קורפוס הארץ. קורפוס זה נבחר בשל איפיונו כבעל צורות בסיס של פועל הרבות ביותר, ועל כן בעל גיוון. בשל השימוש במתייג מורפולוגי שונה, מספר צורות הבסיס של פעלים הינו שונה ממספרם תחת המתייג המורפולוגי של מיל"ה, והוא עומד על 3,288 צורות בסיס, עם 1,337,537 מופעי פועל שונים.

7.1 זיהוי מסגרות ההצרכה

מסגרת ההצרכה שאנו מזהים בחלק זה של עבודתנו זהה להגדרה בה השתמשנו בלמידת מסגרות ההצרכה מקורפוס מתויג מורפולוגית. זוהי מסגרת בת ארגומנט אחד, שהוא צירוף שמני, צירוף יחס עם מילת היחס: אל, ב, ל, מ, על, עם, פסוקית או שם פועל או מסגרת ריקה. ההבדל בין השיטות הוא מיקום המשלימים. בלמידה מקורפוס מתויג מורפולוגית הוספנו אילוץ על מסגרת ההצרכה ודרשנו שהיא תופיעה מיד לאחר הפועל. בלמידה מקורפוס מנותח תחבירית, אנו מרחיבים את קבוצת המשלימים היכולים להתאים לפועל לקבוצת המשלימים המופיעים בקשר עם הפועל. סוגי הקשרים הקיימים במנתח התחבירי זהים לקשרים המופיעים בבנק העצים, ואנו מחפשים משלימים הנמצאים בקשר של OBJ (object), COM (complement) או DEP (dependency) עם הפועל (ראה סעיף 3.2, להסבר על סוגי הקשרים). למרות שהקשר COM הוא קשר חזק יותר מקשר DEP, הרי שיכולת ההפרדה של המנתח אינה מדוייקת כל כך, ועל מנת לקבל תמונה מלאה אנו מרחיבים את הקשרים עליהם אנו מסתכלים, מתוך הנחה כי ההפרדה הנכונה תעשה על ידי המבחנים הסטטיסטיים השונים. בהקשר של שיטה זו, מופע של פועל עם מסגרת הצרכה ריקה הוא מופע שבו לא קיימים משלימים מהקבוצה הנידונה לעיל הנמצאים בקשר OBJ, COM או DEP עם הפועל.

תהליך איסוף המופעים זהה לתהליך בלמידה מקורפוס מתויג מורפולוגית. עבור כל פועל אנו מחשבים ארבעה ערכים החשובים להכרעה בדבר הקשר בין הפועל v למסגרת f:

1. $n_{v,f}$ - מספר מופעי הפועל v עם המסגרת f בקורפוס (כלומר, מספר מופעי הפועל הנמצאים בקשר OBJ, COM או DEP עם מילת יחס, מילת שעבוד, שם פועל או מסגרת ריקה, לפי המסגרת).
2. n_v - מספר מופעי הפועל v בקורפוס.
3. $n_{\neg v,\, f}$ - מספר מופעי פועל כלשהו שאינו v עם המסגרת f (כלומר, מספר מופעי פעלים שונים שאינם v, הנמצאים בקשר OBJ, COM או DEP עם מילת יחס, מילת שעבוד, שם פועל או מסגרת ריקה, לפי המסגרת).
4. $n_{\neg v}$ - מספר מופעי כלל הפעלים השונים מ-v על כל נטיותיהם בזמן, גוף ומין.

זה המקום להעיר כי לעומת המתייג המורפולוגי של מיל"ה, המתייג המורפולוגי שפותח על ידי מני אדלר מספק ניתוח אחד ויחיד ועל כן כל מופע של פועל ומסגרת הצרכה מעלה בדיוק ב-1 את המנייה.

לאחר איסוף הנתונים הללו, אנו משתמשים באותה הטכניקה בה השתמשנו בחלק הלמידה מתוך קורפוס מתויג מורפולוגית לכימות הקשר בין הפועל ומסגרת ההצרכה ולזיהוי ההצרכה. ספי הקבלה בהם אנו משתמשים בחלק זה אף הם זהים לספי הקבלה שנלמדו בחלק הקודם.

7.2 תוצאות

לאחר תהליך הוצאת הזוגות מהקורפורה קיבלנו 21,381 זוגות שונים של פועל-מסגרת הצרכה, עבור 2,955 צורות בסיס של פעלים. לשם השוואה, בתהליך הוצאת הזוגות מקורפורה המתויגים מורפולוגית בלבד, קיבלנו 20,381 זוגות שונים של פועל-מסגרת הצרכה, עבור 3,393 צורות בסיס של פעלים, כך שבממוצע עבור כל פועל בתהליך הנוכחי קיבלנו כ-7.24 משלימים, לעומת 6.14 משלימים בתהליך הקודם. הדבר מראה כי אכן, הסתכלות רחבה יותר סביב הפועל מובילה ליותר משלימים אפשריים לפועל. עבור כל מדד נבנה לקסיקון פועל המכיל זוגות של פועל-מסגרת הצרכה שנלמדו בתהליך, וציונם תחת המדד היה מעל לסף הקבלה של המדד. באופן הזה התקבלו 4 לקסיקוני פועל שונים, השונים זה מזה בפעלים אותם הם כוללים, ובמסגרות ההצרכה המתאימות לכל פועל. נתונים על לקסיקוני הפועל השונים ניתן לראות בטבלה 7.1.

מדד	מספר פעלים בצורת הבסיס	מספר זוגות פועל-מסגרת הצרכה	מספר מסגרות ממוצע לפועל
LLR	427	649	1.52
PMI	2,955	6,545	2.21
RF	2,955	8,180	2.77
t-score	2,955	5,453	1.85

טבלה 7.1 : נתונים על מספר הזוגות פועל-מסגרת הצרכה שהתקבלו תחת המדדים השונים

Table 7.1 : Verb-frames couples data

הסתכלות רחבה יותר סביב הפועל מביאה לידי מציאת מסגרות הצרכה חדשות, שלא יכולנו לקבל בתהליך הקודם, אם בשל העובדה שהמשלים לא הופיע כלל בסמיכות לפועל, ואם בשל מספר מופעים קטן במיקום זה. בסעיף 6.2.7 הוצג הפועל **איחל** כדוגמה לפועל עבורו לא נמצאה מסגרת ההצרכה המלאה הבנויה מהמשלימים צירוף שמני + מילת היחס ׳ל׳. בתהליך הנוכחי, מסגרת ההצרכה הזו נלמדת תחת מדד ה-raw frequency. בתהליך הלמידה מתוך קורפוס מתויג מורפולוגית אחוז מופעי הפועל עם צירוף שמני עמד על כ-9%, ואילו אחוז המופעים עם מילת היחס ׳ל׳ היה 80%. בתהליך הנוכחי, הצירוף השמני מופיע בכ-35% מהמופעים, ומילת היחס ׳ל׳ מופיעה בכ-51% מהמופעים, כך שניתן לראות כי בהחלט הלמידה הנוכחית מציגה תמונה שלמה יותר של מסגרת ההצרכה (ראה טבלה 9.17 בחלק הנספחים) .

כמו כן מציאת משלימים הקשורים לפועל מונעת את ריבוי המקרים של מופעי הצרכה ריקה עבור פעלים בהם הנושא נוטה לבוא לאחר הפועל , כדוגמת הפועל **אמר**. תחת השיטה ללמידת מסגרות הצרכה מתוך קורפוס מתויג מורפולוגית, כ-55% ממופעי הפועל היו מופעים עם מסגרת הצרכה ריקה. תחת השיטה הנוכחית, האחוזים מצטמצמים, ורק כ-33% מהמופעים הם עם מסגרת הצרכה ריקה (טבלה 9.18 בחלק הנספחים).

השוואה בין מספר המשלימים המופיעים כקשורים לפועל (בקורפוס הארץ בלבד) בשתי השיטות השונות מעלה כי, כצפוי, יותר משלימים נחשבים כקשורים לפועל בקורפוס מנותח תחבירית. לעומת זאת, פחות מופעי מסגרת הצרכה ריקה ניצפים, זאת מכיוון שמשלים שהופיע במיקום שאינו סמוך לפועל, ולא נחשב כמשלים בשיטה הראשונה, נלקח עתה בחשבון (טבלה 7.2).

	קורפוס מנותח תחבירית	קורפוס מתויג מורפולוגית
אל	9,156	3,646
את	367,391	131,055
ב	333,745	106,957
ל	190,598	78,418
מ	66,946	21,912
על	76,869	26,140
עם	21,771	1,424
ϕ	326,120	443,349
שם פועל	92,483	58,371
פסוקית	108,871	50,696

טבלה 7.2 : מספר מופעי מסגרות ההצרכה השונות בקורפוס הארץ

Table 7.2 : Ha`aretz Corpus: subcategorization frames

את תוצאות השיטה ללמידת מסגרות ההצרכה מתוך קורפוס מנותח תחבירית אנו מעריכים בעזרת אותם כלים ששימשו להערכת השיטה ללמידת מסגרות ההצרכה מתוך קורופוס מתויג מורפולוגית.

7.2.1 הערכת התוצאות על ידי בחינת קבוצת פעלים בצורה ידנית

הערכת התוצאות נעשית כבפרק 6, על ידי קבוצת פעלים המונה 17 פעלים : **אהב**, **גרם**, **האמין**, **התחיל**, **התלונן**, **התעסק**, **חשב**, **כעס**, **מיהר**, **מצא**, **נראה**, **נתן**, **סיפק**, **עזר**, **ציפה**, **קיבל**, **שאל** (ראה סעיף 6.1).

תוצאות ההערכה (טבלה 7.3) מצביעות על כך שגם בשיטתנו הנוכחית מדד ה-PMI הינו בעל F-score הגבוה ביותר (75.61), אך, לעומת תוצאות הלמידה מקורפוס שהינו מתויג מורפולוגית בלבד, הוא אינו המדד בעל אחוז האחזור (recall) ודיוק (precision) הגבוהים ביותר. המדד בעל אחוז האחזור הגבוה ביותר הינו מדד ה-RF עם אחוז אחזור של 68.75% (לעומת אחוז האחזור הגבוה ביותר תחת שימוש בתיוג מורפולוגי בלבד שעמד על 69.39%), ואילו המדד בעל אחוז הדיוק הגבוה ביותר הינו מדד ה-*t*-score עם 93.75% (לעומת 97.14% בשימוש בתיוג מורפולוגי בלבד). מדד ה-PMI תחת השיטה הנוכחית מורע בכל המדדים, אך דווקא *t*-score ו-RF מצליחים להשתפר במדד האחזור (recall) ומקבלים יותר מסגרות הצרכה נכונות לעומת השימוש בתיוג מורפולוגי בלבד. גם אבחנתו של מדד ה-LLR משתפרת עקב הרחבת מסגרות ההצרכה הפוטנציאליות, ונראה שיפור בכל המדדים. לסיכום, על פי ההערכה הפנימית על קבוצת הפעלים הנוכחית, נראה כי שיטת הלמידה החדשה דווקא אינה מצליחה להביא לשיפור גורף, ואחוזי הדיוק והאחזור המקסימליים תחת שיטה זו עדיין נמוכים לעומת השימוש בתיוג מורפולוגי בלבד (93.75% לעומת 97.14 באחוז הדיוק, ו-68.75% לעומת 69.39% באחוז האחזור).

מדד	TP	TN	FP	FN	Recall %	Precision %	F-score תיוג מורפולוגי	F-score
PMI	31	119	3	17	64.58	91.18	80.95	75.61
***t*-score**	30	120	2	18	62.50	93.75	70.13	75.00
RF	33	102	20	15	68.75	62.26	60.42	65.35
LLR	17	109	12	32	34.69	58.62	39.08	43.59

טבלה 7.3 : תוצאות הערכת קבוצת הפעלים הנבדקת

Table 7.3 : Internal evaluation results

פירוט מסגרות ההצרכה שהתקבלו על ידי המערכת עבור הפעלים השונים, תחת המדדים השונים מופיע בטבלה 7.4 להלן.

פועל	LLR	PMI	RF	*t*-score
אהב	-	את, שם פועל	את, ϕ, שם פועל	את, שם פועל
גרם	ב,ל	ל	את, ל, ϕ	ל
האמין	את, פסוקית	פסוקית	ב, ל, ϕ, פסוקית	פסוקית

פועל	LLR	PMI	RF	*t*-score
התחיל	את, ל, שם פועל	ב, עם, שם פועל	ב, ϕ, שם פועל	ב, שם פועל
התלונן	על	על, פסוקית	ב, על, ϕ, פסוקית	על, פסוקית
התעסק	-	ב, עם	ב, עם	ב, עם
חשב	את, ב, ל, פסוקית	על, פסוקית	על, ϕ, פסוקית	על, פסוקית
כעס	-	על, ϕ	על, ϕ	על, ϕ
מיהר	שם פועל	אל, שם פועל	ϕ, שם פועל	אל, שם פועל
מצא	את	את, ב	את, ב, ל, ϕ	את
נראה	את, ϕ, פסוקית	ϕ, פסוקית	ב, ל, ϕ, פסוקית	ϕ, פסוקית
נתן	את, ל	את, ל	את, ב, ל, ϕ	את, ל
סיפק	את, ל	את, ל	את, ל, ϕ	את, ל
עזר	ל	ל, שם פועל	ב, ל, ϕ, שם פועל	ל, שם פועל
ציפה	ל, פסוקית	ל, מ, פסוקית	ב, ל, פסוקית	ל, מ, פסוקית
קיבל	את, ל, ϕ, שם פועל, פסוקית	את, מ	את, ב, ϕ	את, מ
שאל	-	את, ϕ	את, ב, ϕ	את, ϕ

טבלה 7.4 : תוצאות המערכת עבור קבוצת הפעלים הנבדקים

Table 7.4 : Test verbs results

ייתכן כי ההבדל בין התוצאות בחלק זה לתוצאות החלק הראשון של מחקרנו נעוץ בעבודה כי אנו מרחיבים את אפשרויות הבחירה עבור כל פועל, עם הוספת מסגרות ההצרכה המופיעות רק במרחק ממנו, ולכן אין העדפה גורפת למסגרות הצרכה הנוטות להופיע מיד לאחר הפועל. זהו המקרה למשל בפועל **האמין** בקורפוס PMI. עבור הפועל **האמין** מתקבלת רק מסגרת ההצרכה ′פסוקית′ לעומת מסגרות ההצרכה ′ב′, ′ל′ ו′פסוקית′ המתקבלות בשימוש בתיוג מורפולוגי בלבד. בחינה של הפועל **האמין** מעלה כי המשלימים ′ב′ ו-′ל′ נוטים להופיע מיד לאחר הפועל, לעומת המשלים ′פסוקית′ (בקורפוס הארץ אחוז מופעי ′ב′ ו-′ל′ היה זהה תחת שתי הגישות, לעומת אחוז מופעי ′פסוקית′ שתחת הגישה הנוכחית קפץ לכמעט פי 2). בשל כך הקשר בין הפועל למסגרת ההצרכה ′פסוקית′ התחזק, לעומת הקשר בין הפועל למסגרות ′ב′ ו-′ל′ שנחלש. מלבד קבלת פחות מסגרות הצרכה עבור חלק מהפעלים, בולטת תופעה של קבלת מסגרת הצרכה ריקה, והדבר לעיתים מוביל להגדלת מקרי false positive.

7.2.2 בעיית הצמדת צירוף יחס (pp-attachment)

כבחלק למידת מסגרות ההצרכה מתוך קורופוס מתויג מורפולוגית, גם כאן בחרנו להעריך את התוצאות בבעיית הצמדת צירוף יחס (6.3).

תוצאות הניסויים הושוו לבסיס (baseline) בו קבענו כי צירוף היחס לעולם לא יצטרף לפועל. בחינת הנתונים מראה כי מדד ה-PMI משיג את התוצאות הטובות ביותר (הפחתה של 27.56% בשיעור הטעויות לעומת ה-baseline). בטבלה 7.5 ניתן לראות את תוצאות הניסויים שערכנו עם ארבעת

המדדים LLR, PMI, raw frequency ו-*t*-score, וכן עבור הניסויים Voting ו-OR. בטבלה מופיעים הנתונים על ה-(FN) false negative, (FP) false positive, (TN) true negative ו-(TP) true positive, עבור כל מדד, וכן מדד הדיוק (accuracy), וההפחתה בשיעור הטעות (ERR). לעומת התוצאות שהושגו בלמידת מסגרות ההצרכה מתוך קורפוס מתויג מורפולוגית ניתן לראות שיפור במדד ה-PMI (הפחתה של 8.88% בשיעור הטעויות, לעומת שימוש בתיוג מורפולוגי בלבד) ובמדד ה-LLR (הפחתה של 2.86% בשיעור הטעויות, לעומת שימוש בתיוג מורפולוגי בלבד), אך במדדים raw frequency ו-*t*-score לא נראה שיפור. את חוסר השיפור במדד ה-raw frequency ניתן להסביר בכך שמכיוון שבשיטה זו ישנם יותר משלימים המופיעים עם הפועל, ומנגד, קיימים משלימים מוצרכים המופיעים בעיקר מיד לאחר הפועל, יוצא שהוספת רעש (=משלימים שאינם מוצרכים לפועל) מפחיתה את אחוז מופעי המשלימים המקוריים, ובשל כך בחלק מהמקרים ההשלמה לא מאותרת.

Method	FN	FP	TN	TP	Accuracy	Accuracy תיוג מורפולוגי	ERR	ERR לעומת תיוג מורפולוגי
Baseline_FALSE	156	0	167	0	51.70%			
LLR	108	27	140	48	58.20%	56.97%	13.46%	2.86%
PMI	**58**	**55**	**112**	**98**	**65.02%**	**61.61%**	**27.56%**	**8.88%**
RF	37	89	78	119	60.99%	62.54%	19.23%	-4.14%
t-score	72	44	123	84	64.09%	65.63%	25.64%	-4.48%
Voting	79	37	130	77	64.09%	63.47%	25.64%	1.70%
OR	21	108	59	135	60.06%	59.75%	17.31%	0.77%

טבלה 7.5 : תוצאות הצמדת מילות היחס

Table 7.5 : PP-attachment results

עבור הצמדת המיליות ′ש′ או ′כי′, ניתן לראות (טבלה 7.7) שיפור במדד ה-PMI ומדד ה-*t*-score (הפחתה של 50% בשיעור הטעויות, לעומת שימוש בתיוג מורפולוגי בלבד).

Method	FN	FP	TN	TP	Accuracy	Accuracy תיוג מורפולוגי	ERR	ERR לעומת תיוג מורפולוגי
Baseline_FALSE	4	0	16	0	80.00%			
LLR	0	4	12	4	80.00%	80.00%	0.00%	0.00%
PMI	**0**	**1**	**15**	**4**	**95.00%**	**90.00%**	**75.00%**	**50.00%**
RF	**0**	**1**	**15**	**4**	**95.00%**	**95.00%**	**75.00%**	**0.00%**
***t*-score**	**0**	**1**	**15**	**4**	**95.00%**	**90.00%**	**75.00%**	**50.00%**
Voting	**0**	**1**	**15**	**4**	**95.00%**	**95.00%**	**75.00%**	**0.00%**
OR	0	4	12	4	80.00%	75.00%	0.00%	20.00%

טבלה 7.6 : תוצאות הצמדת הפסוקיות

Table 7.6 : Complementizer-attachment results

7.3 סיכום

למרות היתרון בשימוש בקורפוס מנותח תחבירית, ייתכן כי ריבוי המשלימים הקשורים לפועל, אף שאינם מוצרכים לו, הוא הגורם לכך שקשה יותר לסווג את אותם משלימים המוצרכים לפועל, ואת אלו שהם רק בבחינת משלימים תיאוריים. משלימים מוצרכים הנוטים להופיע מיד אחרי הפועל, ובשל ייחודיות זו הקשר בין המשלים לפועל הצליח להילמד בשימוש בתיוג מורפולוגי בלבד, מאבדים את הקשר הייחודי הזה בטכניקה הנוכחית, ואובדים בתוך שלל המשלימים המופיעים סביב הפועל. מנגד, השיטה מאפשרת ללמוד מסגרות הצרכה הנוטות להופיע רחוק מהפועל, דבר שלא התאפשר כלל בגישתנו הראשונה. יכול להיות ששימוש בטכניקה של Sarkar ו-Zeman (2000), כאשר מנסים ללמוד את מסגרת ההצרכה המלאה, ולא החלקית, תוכל להניב תוצאות טובות יותר, זאת מכיוון שמסגרת ההצרכה הרחבה היא קבועה, ובתהליך החיפוש אחר מסגרת הצרכה רחבה ויציבה, נוכל להבחין במשלימים התיאוריים. כמו כן ייתכן כי שילוב שתי השיטות יניב תוצאות מדוייקות יותר.

פרק 8

סיכום ומחקר עתידי

מטרת מחקרנו הייתה יצירת לקסיקון פועלי מקיף שיכיל את מסגרת ההצרכה השלמה של הפועל. לאור המשאבים המועטים העומדים לרשותנו בעברית, ומורכבות השפה והמשימה, נאלצנו להצטמצם לכדי מציאת מסגרות הצרכה חלקיות בלבד. למרות העובדה כי מצאנו רק מסגרות הצרכה חלקיות, ניתן לראות כי אף הן יכולות להוות כלי חשוב ומרכזי לפתרון בעיות התלויות באופן ישיר במסגרות ההצרכה, כדוגמת הצטרפות מילות היחס (סעיף 6.2) ותרגום אוטומטי של פעלים ומסגרות ההצרכה שלהם (6.3).

מחקרנו התרכז במציאת מסגרות ההצרכה החלקיות של הפועל, ובלקסיקון הפועל שיצרנו מופיעות מספר מסגרות הצרכה חלקיות לכל פועל (בממוצע 2.1 מסגרות הצרכה חלקיות לפועל תחת מדד ה-PMI, לדוגמה). כמו שציינו בפרק 1, בסקירת התיאוריה הבלשנית מאחורי מסגרות ההצרכה, לפועל אמנם תיתכנה מספר מסגרות הצרכה שונות, אך ישנם פעלים הזקוקים למסגרת הצרכה רחבה, בת יותר ממשלים אחד, ולכן ייתכן כי מסגרות ההצרכה החלקיות מהוות רכיבים שונים של אותה מסגרת הצרכה רחבה יותר. הסיבה לקבלת מספר רכיבים שונים של אותה מסגרת ההצרכה, למרות שהתרכזנו במשלימים המופיעים מיד לאחר הפועל, נעוצה בגמישות היחסית בסדר המשלימים במסגרת ההצרכה. מחקר עתידי יוכל להשתמש בנתוני לקסיקון הפועל החלקיים שמצאנו, וליצור האחדה בין מסגרות חלקיות שונות.

את המחקר ניתן אף להרחיב בשימוש במסגרות ההצרכה לשיפור הפגת העמימות המורפולוגית. כמו כן ניתן אף להשתמש במידע לשיפור המנתח התחבירי, כיוון שהוספת מידע על דרישות הפועל יכולה לשפר את ההחלטות בדבר הקשרים במשפט.

מחקר עתידי יכול להתרכז אף במציאת צירופים פועליים כבולים, כדוגמת: *הביא בחשבון* או *התקבל על הדעת*. בצירופים שכאלה קיים קשר של הצרכה כפולה. מלבד קשר ההצרכה (הייחודי לעיתים לצירוף) בין הפועל למילת היחס, יש קשר של הצרכה בין מילת היחס לשם העצם. כך, כאשר בוחנים את הפועל

הביא, הוא איננו דורש את מילת היחס **ב**, אולם תחת הצירוף הכבול ישנו קשר של הצרכה. מציאת צירופים פועליים כבולים יכולה אף היא לשפר הן את התיוג המורפולוגי, והן את הניתוח התחבירי.

מלבד הפועל תיתכן גם הצרכה לחלקי דיבר נוספים, כדוגמת שמות עצם. מחקר עתידי יוכל להרחיב את מציאת מסגרות ההצרכה גם עבורם.

נספחים

נספח א: מסגרות ההצרכה של הפעלים השונים

הפעלים הבאים הינם פעלים שהופיעו בחיבור בפרקים 6 ו-7. עבור כל פועל מובאות מסגרות ההצרכה, על פי בחינה שנעשתה באופן מדוקדק על כל פועל ופועל על ידי בוחן דובר עברית. עבור כל מסגרת הצרכה מובאת דוגמה מתוך טקסטים במשלב העיתונאי.

הפועל 'אהב'

את: סבא **אהב את** משפחתו, אהב את אשתו, אהב את העבודה. (הארץ, 31.3.02)

שם פועל: הקיסר דומיטיאנוס **אהב לצוד** זבובים ולהוקיעם על מוט מחודד. (הארץ, 11.8.06)

פסוקית: הוא לא **אוהב ש**'דופקים' אותו". (הארץ, 28.2.07)

הפועל 'איחל'

את: דה וילפן **איחל חג שמח** לעובדי שגרירות ישראל. (הארץ, 22.12.05)

ל: ראש הממשלה **איחל לרמטכ"ל** ולצה"ל שנה טובה. (ידיעות אחרונות, 27.9.11)

פסוקית: למוסדיים הוא **מאחל ש**-2006 לא תהיה יותר גרועה מ-2005, ושיפעילו עוצמה אפקטיווית בשוק ההון. (דה מרקר, 27.12.05)

הפועל 'אילץ'

את: משרד התקשורת **אילץ את בזק** לבטל הסכם עם משרד הביטחון. (דה מרקר, 25.4.07)

שם פועל: המיתון הכלכלי **אילץ** חברות **לפטר** עובדים. (הארץ, 4.12.01)

הפועל 'אמר'

את: נאמן **אמר את הדברים** בכנס רבנים ודיינים שנערך אמש. (הארץ, 8.12.09)

ל: חבל שג'ינו לא בא **ואמר לי** את זה בארבע עיניים. (הארץ, 21.7.05)

על: עורך הדין **אמר על** פיוטרקובסקי: "הוא לא רק שקרן, הוא מעבר לזה". (דה מרקר, 22.6.06)

פסוקית: גלייזר **אמר כי** לא יגביל את סכומי ההשקעה בשחקנים חדשים. (דה מרקר, 3.7.05)

הפועל 'גרם'

את: שתי תאונות בין רכבים למטוסים בנתב"ג **גרמו נזק** של עשרות אלפי דולרים. (דה מרקר, 28.4.05)

ל: הגשמים **גרמו לנ**זקים של מיליוני שקלים לחקלאי הערבה. (דה מרקר, 28.2.10)

שם פועל: בית המשפט **גרם** לו **לחשוב** שהוא יכול להמשיך להתנהל בחוסר זהירות. (דה מרקר, 8.5.12)

הפועל 'האמין'

ב: אדריכל האמריקאי **האמין ב**ארכיטקטורה אורגנית. (הארץ, 6.6.07)

ל: שני שלישים מהציבור **מאמינים ל**"כלבוטק". (הארץ, 16.10.07)

על: "שכונות סגורות" על פני המים? מי היה **מאמין על** אמסטרדם. (הארץ, 2.8.07)

פסוקית: ברלוסקוני אמר אתמול כי אינו **מאמין ש**יידרשו יותר מארבעה שבועות להחלטה על פעולה צבאית. (הארץ, 4.2.03)

הפועל 'הביא'

את: האסון **הביא את הצופים** אמש למסך. (הארץ, 3.12.10)

ל: חוסר הסכמה בין חברי הקואליציה **הביא ל**דחיית ההצבעה. (דה מרקר, 31.10.10)

הפועל 'הועמד'

ל: מגדל "המלפפון החמוץ" בלונדון **הועמד ל**מכירה ב-1.1 מיליארד דולר. (דה מרקר, 18.9.06)

על: המחיר שקיבלה **הועמד על** 100 אלף דולר לדונם. (הארץ, 26.2.02)

הפועל 'החליט'

על: ביהמ"ש המחוזי בת"א **החליט על** ביטול הליכי הפירוק. (דה מרקר, 8.7.03)

שם פועל: ליברמן **החליט להאריך** את קו הרכבת הקלה בירושלים. (הארץ, 21.5.03)

פסוקית: רוכש מקבצי הדיור **החליט ש**לא לממש את העסקה. (הארץ, 31.5.05)

הפועל 'התגורר'

ב: הוא **התגורר ב**חולון עם אביו משה. (הארץ, 16.2.02)

הפועל 'התחיל'

את: ריבלין **התחיל את** הקריירה השיפוטית שלו כשופט תעבורה ב-1976. (הארץ, 28.5.12)

ב: מרתון הספינינג **התחיל ב**-09:30 בבוקר. (הארץ, 1.5.05)

עם: זה **התחיל עם** הצלת בר סטרנס רגע לפני שפשט רגל. (הארץ, 12.10.08)

שם פועל: מנהל המוזיאון **התחיל** השבוע **לשרוף** יצירות אמנות של המוזיאון. (הארץ, 19.4.12)

הפועל 'התייעץ'

על: השר לא **התייעץ על** כך עם ראש עיריית ירושלים. (הארץ, 29.8.01)

עם: צור המשיך **להתייעץ עם** מנהל האצטדיון. (הארץ, 3.12.01)

הפועל 'התלונן'

ל: עובדים בחברה **התלוננו ל**-TheMarker על "התנהלות לא תקינה". (דה מרקר, 11.12.07)

על: ייני **התלונן על** כאבים בכתף. (הארץ, 11.9.08)

פסוקית: בוחן אחר **התלונן ש**המרחב מלפנים צפוף במקצת. (דה מרקר, 4.12.07)

הפועל 'התנפל'

על: כלב מגזע פיט בול **התנפל על** צעירה בת 24. (הארץ, 16.8.02)

הפועל 'התעסק'

ב: שלומי זוכר שעד גיל מאוחר לא **התעסק ב**לבושו. (הארץ, 6.4.12)

עם: כל העולם **התעסק עם** המשימה המסובכת הזו ובסוף נכשלנו. (דה מרקר, 25.7.06)

הפועל 'חשב'

על: רוברטסון מספר כי מיד **חשב על** המחווה המפורסמת. (הארץ, 6.4.12)

שם פועל: מי ש**חשב לתת** למחזיקי האג"ח נזיד עדשים – טעה. (הארץ, 19.4.12)

פסוקית: רה"מ בטח **חשב ש**הוא מגיע לעוד נאום שגרתי. (דה מרקר, 5.11.07)

הפועל 'כעס'

על: גולשים רבים **כעסו על** הקמפיין של לאומי. (הארץ, 9.1.12)

ϕ: המטרה היא שיהיה "הפי אנד" והציבור יהיה מרוצה. שלא **יכעס**, חס וחלילה. (דה מרקר, 4.4.12)

הפועל 'מחה'

את: היטלר **מחה את הצבא** היחידי שיכול היה להתנגד לו בין ברלין לדנובה. (הארץ, 16.3.09)

על: נשיא האיגוד העולמי של מועצות העיתונות **מחה על** הפגיעה בחופש העיתונות בשטחים. (הארץ, 9.4.12)

הפועל 'מיהר'

אל: הוא **מיהר אל** מסך הטלוויזיה. (הארץ, 20.1.12)

שם פועל: שמעון פרס **מיהר לחגוג** את נצחונו. (הארץ, 18.4.12)

הפועל 'מצא'

את: שוורצקופף **מצא שותפים**. (הארץ, 18.4.05)

פסוקית: סקר **מצא ש**-11% מבין 200 הראשונים בעולם היו שמאליים. (הארץ, 30.1.11)

הפועל 'נאלץ'

שם פועל: דיוויד בואי **נאלץ להפסיק** הופעה בשל דלקת בכתף. (הארץ, 28.6.04)

הפועל 'נדבק'

אל: מקטע התוכנה **נדבק אל** מקטע הקלט, ואנזים מחבר אותם יחד. (הארץ, 22.11.01)

ב: תושב ניו יורק **נדבק ב**נגיף ה-HIV כתוצאה מתרומת כליה. (הארץ, 17.3.11)

ל: דניס **נדבק ל**"פחדרון בארון". (הארץ, 12.6.03)

מ: התינוק **נדבק מ**המוהל בהרפס ונותר עם פגיעה מוחית קשה. (ידיעות אחרונות, 8.10.02)

הפועל 'נהר'

אל: אלפי גברים, מבוגרים ונערים **נהרו אל** האור הבוהק. (הארץ, 9.11.11)

ל: אלפי חובבי בירה מהעולם **נהרו** אתמול **ל**מינכן. (הארץ, 21.9.03)

הפועל 'נזקק'

ל: העולם **נזקק ל**פעולות קולקטיוויות לטיפול במשבר. (הארץ, 21.8.09)

הפועל 'ניצֵל'

את: מוטי זיסר **ניצל את** הנפילות בשוק. (דה מרקר, 16.9.07)

הפועל 'ניצַל'

מ: היכל התרבות **ניצל מ**"שיפוצים" בנוסח הבימה. (הארץ, 28.8.09)

הפועל 'נמצא'

ב: חיידק עמיד **נמצא ב**-3 תינוקות בפגיה בבי"ח רמב"ם בחיפה. (גלובס, 7.3.12)

ϕ: **נמצא** פתרון לבעיות החניה בתל אביב. (דה מרקר, 24.7.11)

פסוקית: במעקב **נמצא כי** נבדקים אלו היו בסיכון גבוה פי 3.6 לתמותה. (הארץ, 5.4.12)

הפועל 'נראה'

ב: גיא **נראה** לאחרונה **ב**בסיס בו שירת ברמת הגולן. (הארץ, 17.8.08)

ל: מחיר של 10 דולרים **נראה ל**אנליסטים של CIBC כסביר בהחלט. (דה מרקר, 13.6.06)

פסוקית: **נראה ש**-10 המדינות המזרחיות לא יוכלו להסתמך על התאוששות הביקוש המקומי. (דה מרקר, 21.5.09)

הפועל 'נתן'

את: ליצמן **נתן את האישור** לפיילוט מבלי שהתוכנית אושרה על ידי הדרג המקצועי במשרד. (הארץ, 18.4.12)

ל: לא ראיתי מי **נתן לי** את המכה. (הארץ, 19.4.12)

שם פועל: הוציא כסף ובקבוק שנאפס מהכיס **ונתן** לכולם **לשתות**. (הארץ, 25.5.12)

הפועל 'סיפק'

את: מפעל ברמד **סיפק ציוד** לבית זיקוק גדול במקסיקו בכ-2.5 מיליון שקל. (דה מרקר, 6.4.03)

ל: דאפר **סיפק ל**פעילים של החזית הדמוקרטית כלי נשק. (הארץ, 6.7.03)

הפועל 'עזר'

ב: ראש לשכת ראש העיר הסביר שבן לולו **עוזר ב**"סגירות האחרונות" מול משרדי הממשלה. (הארץ, 24.8.05)

ל: אנשי תעאיוש **עזרו ל**דרי המערות בקציר. (הארץ, 5.5.02)

שם פועל: מחירי הנפט הגבוהים **עזרו לרשום** רווח של 5.6 מילירד דולר ברבעון השני. (דה מרקר, 26.7.11)

הפועל 'פונה'

אל: בית המעצר אבו כביר **יפונה אל** מחוץ העיר לטובת שימושי תעסוקה. (דה מרקר, 18.12.11)

ל: פסולת רעילה תעשיתית **מפונה ל**אתר הפסולת הארצי ברמת חובב. (הארץ, 30.9.03)

מ: מתבצרים **מפונים מ**"בית שפירא" בעקבות הוראת בג"ץ. (הארץ, 16.4.08)

הפועל 'ציפה'

ל: השוק **ציפה ל**תוצאות טובות יותר. (דה מרקר, 24.11.10)

מ: סולברג **ציפה מ**"עובדה" לדיוק עובדתי כמו בית משפט. (הארץ, 9.2.12)

שם פועל: הזוג **ציפה לגייס** מיליון דולר לקרן הצדקה של שרון אוסבורן. (דה מרקר, 2.12.07)

פסוקית: דניאל **ציפה ש**הישגיו יעניקו לו מעמד של ספורטאי מצטיין. (הארץ, 19.2.12)

הפועל 'קיבל'

את: לנקרי **קיבל קביעות**. (הארץ, 17.9.07)

מ: הוא אומנם **קיבל מ**-Ethiopian Airlines 150 דולר. (הארץ, 26.7.05)

על: הבונוס הגיע לסך כולל של פי כמה מהסכום **שקיבלו על** העבודה עצמה. (דה מרקר, 4.5.12)

הפועל 'רצה'

את: גרשון **רצה את מאציאוסקאס** ופן במכבי. (הארץ, 23.7.06)

שם פועל: מי באמת **רצה להרוס** את גימנסיה הרצליה? (הארץ, 8.5.09)

פסוקית: ספילברג **רצה ש**"סוס מלחמה" יהיה מעין אלגוריה אנטי מלחמתית. (הארץ, 25.1.12)

הפועל 'שאל'

את: החוקרים **שאלו את המרואיינים** כיצד הם מכנים כלי מטבח. (הארץ, 12.3.12)

על: קרלוס **שאל על** חומר הנפץ הפלסטי סי-4. (הארץ, 27.4.12)

הפועל 'תם'

ϕ: **תם** עידן האותיות הקטנות. (הארץ, 18.4.12)

נספח ב: טבלאות פרק 6

בטבלאות הבאות נשתמש במוסכמות הללו:

- מסגרות ההצרכה שאמורות היו להתקבל (על פי הניתוח הידני בחלק הנספחים) יודגשו על ידי ***מסגרת הצרכה*.**
- מקרי true positive - מסגרות שהתקבלו תחת מדד מסויים, ואכן היו אמורות להתקבל, יסומנו ב- (**TP**) .
- מקרי false positive - מסגרות שהתקבלו תחת מדד מסויים, אך לא היו אמורות להתקבל, יסומנו ב- (**FP**) .

verb-frame	RF	LLR	*t*-score	PMI
תם+אל	-	-	-	-
תם +את	0.0066	158.1687	-0.2928	-3.0979
תם +ב	0.0082	102.2247	-0.2383	-2.5721
תם +ל	0.0016	111.9326	-0.2368	-4.0631
תם +מ	-	-	-	-
תם +על	0.0001	41.2277	-0.1384	-6.1014
תם +עם	0.0016	1.8593	-0.0352	-1.1108
תם	0.9802 (**TP**)	829.5580 (**TP**)	0.7777 (**TP**)	0.7611 (**TP**)
תם +שם פועל	-	-	-	-
תם +פסוקית	0.0017	68.1365	-0.1850	-3.6186

טבלה 9.1 : הפועל 'תם'

Table 9.1 : '*tm*'

verb-frame	RF	LLR	*t*-score	PMI
אמר+אל	0.0002	455.0119	-0.0461	-2.8458
אמר+את	0.0556	6081.0904 (**TP**)	-0.1976	-0.9641
אמר+ב	0.0477	3435.9068 (**FP**)	-0.1500	-0.8172
אמר+ל	0.1070	112.0960	0.0305	0.1147 (**TP**)
אמר+מ	0.0012	2538.1708 (**FP**)	-0.1090	-2.9099
אמר+על	0.0062	2570.1268 (**TP**)	-0.1180	-1.7016
אמר+עם	0.0004	527.4922	-0.0503	-2.5453
אמר	0.5491 (**FP**)	2480.2435 (**FP**)	0.1414 (**FP**)	0.1817 (**FP**)
אמר+שם פועל	0.0031	8205.6438 (**FP**)	-0.1969	-3.0665
אמר+פסוקית	0.2295 (**TP**)	22871.5285 (**TP**)	0.5441 (**TP**)	1.3145 (**TP**)

טבלה 9.2 : הפועל 'אמר'

Table 9.2 : '*amr*'

PMI	*t*-score	LLR	RF	verb-frame
-	-	-	-	החליט+אל
-2.0508	-0.2687	(**FP**) 1986.6353	0.0188	החליט+את
-0.9566	-0.1597	(**FP**) 594.3634	0.0415	החליט+ב
-2.8133	-0.2277	(**FP**) 1583.3235	0.0057	החליט+ל
-2.6892	-0.1037	336.2604	0.0015	החליט+מ
(**TP**) 0.6766	(**TP**) 0.1348	262.0955	0.0664	***החליט+על***
-2.1297	-0.0465	63.8089	0.0006	החליט+עם
-0.1047	-0.0681	85.5581	(**FP**) 0.4124	החליט
(**TP**) 1.6765	(**TP**) 0.8694	(**TP**) 7335.0029	(**TP**) 0.3564	***החליט+שם פועל***
(**TP**) 0.4512	0.1090	187.3803	0.0968	***החליט+פסוקית***

טבלה 9.3 : הפועל 'החליט'

Table 9.3 : '*hxliT*'

PMI	*t*-score	LLR	RF	verb-frame
(**TP**) 0.9952	0.0802	7.3463	0.0108	***נהר+אל***
-2.1007	-0.2691	184.3401	0.0179	נהר+את
-2.2044	-0.2296	138.1927	0.0119	נהר+ב
-1.0454	-0.1562	53.3846	0.0335	**נהר+ל**
-1.6164	-0.0887	19.6410	0.0043	נהר+מ
-3.4406	-0.1343	54.5557	0.0011	נהר+על
-	-	-	-	נהר+עם
(**FP**) 0.6876	(**FP**) 0.6744	(**FP**) 860.0481	(**FP**) 0.9107	נהר
-1.9232	-0.1693	73.7198	0.0097	נהר+שם פועל
-	-	-	-	נהר+פסוקית

טבלה 9.4 : הפועל 'נהר'

Table 9.4 : '*nhr*'

PMI	*t*-score	LLR	RF	verb-frame
(**TP**) 3.5005	(**TP**) 1.5126	(**TP**) 698.1123	(**TP**) 0.1325	***פונה+אל***
-2.0152	-0.2658	194.6203	0.0195	פונה+את
-0.4534	-0.0941	18.4085	0.0686	פונה+ב
(**TP**) 1.0833	(**TP**) 0.4710	282.4929	(**TP**) 0.2819	***פונה+ל***
(**TP**) 1.1219	(**TP**) 0.2293	62.7635	0.0669	***פונה +מ***
-1.8566	-0.1170	38.6734	0.0053	פונה+על
-0.9238	-0.0317	2.3901	0.0020	פונה +עם
-0.1029	-0.0667	8.2189	(**FP**) 0.4131	פונה
-1.9768	-0.1708	82.7318	0.0092	פונה+שם פועל
-4.1323	-0.1871	118.2992	0.0010	פונה +פסוקית

טבלה 9.5 : הפועל 'פונה'

Table 9.5 : '*pnh*'

verb-frame	RF	LLR	*t*-score	PMI
נדבק+אל	0.0741 (**TP**)	30.7889	0.8239 (**TP**)	2.9194 (**TP**)
נדבק+את	0.0322	14.9521	-0.2388	-1.5103
נדבק+ב	0.3454 (**TP**)	41.2099	0.5676 (**TP**)	1.1634 (**TP**)
נדבק+ל	0.1870 (**TP**)	8.0745	0.2311 (**TP**)	0.6728 (**TP**)
נדבק+מ	0.0355	0.7625	0.0694	0.4865 (**TP**)
נדבק+על	-	-	-	-
נדבק+עם	-	-	-	-
נדבק	0.3032 (**FP**)	10.3229	-0.2302	-0.4121
נדבק+שם פועל	0.0226	4.2771	-0.1311	-1.0833
נדבק+פסוקית	-	-	-	-

טבלה 9.6 : הפועל 'נדבק'

Table 9.6 : '*ndbq*'

verb-frame	RF	LLR	*t*-score	PMI
נמצא+אל	-	-	-	-
נמצא+את	0.0263	1,994.3606 (**FP**)	-0.2531	-1.7151
נמצא+ב	0.3635 (**TP**)	5,495.4485 (**TP**)	0.6162 (**TP**)	1.2144 (**TP**)
נמצא+ל	0.0123	1,484.6592 (**FP**)	-0.2111	-2.0457
נמצא+מ	0.0045	250.8838	-0.0886	-1.5846
נמצא+על	0.0164	136.6120	-0.0717	-0.7204
נמצא+עם	0.0097	42.3151	0.0499	0.6648 (**FP**)
נמצא	0.5529 (**TP**)	437.2877	0.1423 (**TP**)	0.1885 (**TP**)
נמצא+שם פועל	0.0008	1,556.1671 (**FP**)	-0.1971	-4.4159
נמצא+פסוקית	0.0137	689.6658 (**TP**)	-0.1489	-1.5057

טבלה 9.7 : הפועל 'נמצא'

Table 9.7 : '*nmca*'

verb-frame	RF	LLR	*t*-score	PMI
התגורר+אל	-	-	-	-
התגורר+את	0.0283	208.8439	-0.2472	-1.6387
התגורר+ב	0.5646 (**TP**)	1,648.0065 (**TP**)	1.0931 (**TP**)	1.6548 (**TP**)
התגורר+ל	0.0045	215.7007	-0.2295	-3.0464
התגורר+מ	0.0053	24.2314	-0.0839	-1.4157
התגורר+על	0.0030	63.3654	-0.1263	-2.4133
התגורר+עם	0.0083	2.4657	0.0351	0.5120 (**FP**)
התגורר	0.3821 (**FP**)	31.0351	-0.1129	-0.1810
התגורר+שם פועל	-	-	-	-
התגורר+פסוקית	0.0038	129.8775	-0.1785	-2.7919

טבלה 9.8 : הפועל 'התגורר'

Table 9.8 : '*htgwrr*'

PMI	*t*-score	LLR	RF	verb-frame
-	-	-	-	הועמד+אל
-2.0090	-0.2656	194.0951	0.0196	הועמד+את
-0.4156	-0.0877	15.8681	0.0712	הועמד+ב
(**TP**) 1.7444	(**TP**) 1.1386	(**TP**) 1294.2614	(**TP**) 0.5460	***הועמד+ל***
-1.4832	-0.0856	19.5389	0.0049	הועמד+מ
(**TP**) 0.1330	0.0197	0.6846	0.0386	***הועמד+על***
-	-	-	-	הועמד+עם
-0.5646	-0.2942	167.5524	(**FP**) 0.2603	הועמד
-0.3186	-0.0541	5.9056	0.0485	הועמד+שם פועל
-1.9351	-0.1627	74.7844	0.0089	הועמד+פסוקית

טבלה 9.9 : הפועל 'הועמד'

Table 9.9 : '*hw'md*'

PMI	*t*-score	LLR	RF	verb-frame
-0.6568	-0.0228	13.4336	0.0021	הביא+אל
(**TP**) 0.5805	(**TP**) 0.2428	(**TP**) 1,062.3194	(**TP**) 0.2608	***הביא+את***
-0.9072	-0.1548	(**FP**) 643.5767	0.0436	הביא+ב
(**TP**) 1.2547	(**TP**) 0.6089	(**TP**) 5,141.5397	(**TP**) 0.3346	***הביא+ל***
-1.4156	-0.0844	217.1851	0.0053	הביא+מ
-0.8767	-0.0815	180.3646	0.0141	הביא+על
(**FP**) 1.4262	(**FP**) 0.1671	333.4203	0.0207	הביא+עם
-0.3998	-0.2261	(**FP**) 1,127.7228	(**FP**) 0.3070	הביא
-1.9106	-0.1700	(**FP**) 940.2947	0.0099	הביא+שם פועל
-3.4020	-0.1850	(**FP**) 1,291.2310	0.0021	הביא+פסוקית

טבלה 9.10 : הפועל 'הביא'

Table 9.10 : '*hbia*'

PMI	*t*-score	LLR	RF	verb-frame
-	-	-	-	איחל+אל
-0.49902	-0.12045	15.12706	0.088601	***איחל+את***
-2.62294	-0.23924	85.64529	0.007833	איחל+ב
(**TP**) 2.127861	(**TP**) 1.78324	(**TP**) 1415.687	(**TP**) 0.801198	***איחל+ל***
-2.39658	-0.10064	15.38144	0.001984	איחל+מ
-	-	-	-	איחל+על
-	-	-	-	איחל+עם
-1.86888	-0.57652	371.9353	0.070651	איחל
-3.51432	-0.1923	60.3746	0.001984	איחל+שם פועל
-0.79819	-0.10453	12.45383	0.027749	***איחל+פסוקית***

טבלה 9.11 : הפועל 'איחל'

Table 9.11 : '*aixl*'

verb-frame	RF	LLR	*t*-score	PMI
אילץ+אל	-	-	-	-
אילץ+את	(**TP**) 0.4939	(**TP**) 593.9556	(**TP**) 0.7314	(**TP**) 1.2192
אילץ+ב	0.0057	160.3914	-0.2444	-2.9436
אילץ+ל	0.0008	168.5491	-0.2390	-4.8354
אילץ+מ	0.0011	30.8280	-0.1050	-2.9535
אילץ+על	-	-	-	-
אילץ+עם	0.0011	3.8247	-0.0405	-1.4777
אילץ	(**FP**) 0.1512	374.8318	-0.4566	-1.1078
אילץ+שם פועל	(**TP**) 0.3457	(**TP**) 592.8963	(**TP**) 0.8303	(**TP**) 1.6462
אילץ+פסוקית	0.0004	107.8615	-0.1890	-5.0919

טבלה 9.12 : הפועל 'אילץ'

Table 9.12 : '*ailc*'

verb-frame	RF	LLR	*t*-score	PMI
נאלץ+אל	-	-	-	-
נאלץ+את	0.0179	535.3932	-0.2693	-2.0984
נאלץ+ב	0.0119	401.5308	-0.2298	-2.2030
נאלץ+ל	0.0042	443.1952	-0.2305	-3.1179
נאלץ+מ	0.0011	94.4399	-0.1052	-2.9706
נאלץ+על	-	-	-	-
נאלץ+עם	-	-	-	-
נאלץ	(**FP**) 0.2252	(**FP**) 628.6104	-0.3469	-0.7098
נאלץ+שם פועל	(**TP**) 0.7383	(**TP**) 7,774.7124	(**TP**) 2.0105	(**TP**) 2.4049
נאלץ+פסוקית	0.0014	306.1154	-0.1861	-3.8096

טבלה 9.13 : הפועל 'נאלץ'

Table 9.13 : '*nalc*'

verb-frame	RF	LLR	*t*-score	PMI
ניצל+אל	-	-	-	-
ניצל+את	(**TP**) 0.73556	(**TP**) 1825.133	(**TP**) 1.240086	(**TP**) 1.617476
ניצל+ב	0.023306	109.4468	-0.20232	-1.53264
ניצל+ל	0.008417	146.4296	-0.21971	-2.42801
ניצל+מ	0.008721	10.72973	-6.64E-02	-0.91605
ניצל+על	0.000387	66.81969	-0.13712	-4.4692
ניצל+עם	0.001908	2.587661	-3.24E-02	-0.95988
ניצל	(**FP**) 0.220394	253.1994	-0.35368	-0.73121
ניצל+שם פועל	2.18E-05	142.6051	-0.1982	-8.02503
ניצל+פסוקית	0.001286	118.7705	-0.1862	-3.86949

טבלה 9.14 : הפועל 'ניצֵל' (בניין פיעל)

Table 9.14 : '*nicl*' (Pi`l)

PMI	*t*-score	LLR	RF	verb-frame
-	-	-	-	ניצל+אל
(**FP**) 1.2031	(**FP**) 0.7147	523.2356	(**FP**) 0.4860	ניצל+את
-0.7134	-0.1316	30.6761	0.0529	ניצל+ב
-2.5079	-0.2213	116.7857	0.0078	ניצל+ל
(**TP**) 0.8905	(**TP**) 0.1591	26.6553	0.0531	***ניצל+מ***
-2.0275	-0.1204	33.5328	0.0044	ניצל+על
-5.1867	-0.0522	7.7926	0.0000	ניצל+עם
-0.1509	-0.0955	13.5329	(**FP**) 0.3937	ניצל
-4.0088	-0.1946	101.5296	0.0012	ניצל+שם פועל
-4.3133	-0.1876	95.6422	0.0008	ניצל+פסוקית

טבלה 9.15 : הפועל 'ניצַל' (בניין נפעל)

Table 9.15 : '*nicl*' (Nf`l)

PMI	*t*-score	LLR	RF	verb-frame
-	-	-	-	מחה+אל
-1.6269	-0.2463	84.8729	0.0287	***מחה+את***
-0.6905	-0.1286	19.5672	0.0541	מחה+ב
-	-	-	-	מחה+ל
-1.6218	-0.0888	11.5587	0.0043	מחה+מ
(**TP**) 2.3400	(**TP**) 1.3014	(**TP**) 609.9486	(**TP**) 0.3506	***מחה+על***
-	-	-	-	מחה+עם
(**FP**) 0.1789	(**FP**) 0.1335	17.4780	(**FP**) 0.5476	מחה
-2.4882	-0.1817	53.4398	0.0055	מחה+שם פועל
-1.8993	-0.1617	39.3839	0.0092	מחה+פסוקית

טבלה 9.16 : הפועל 'מחה'

Table 9.16 :'*mxh*'

נספח ג: טבלאות פרק 7

verb-frame	RF	LLR	*t*-score	PMI
איחל+אל	-	-	-	-
איחל+את	0.3511 (**TP**)	2.6182	0.1480	0.2454
איחל+ב	0.1064	12.2780	-0.2860	-0.8525
איחל+ל	0.5106 (**TP**)	70.9299	0.9107 (**TP**)	1.2763 (**TP**)
איחל+מ	-	-	-	-
איחל+על	-	-	-	-
איחל+עם	-	-	-	-
איחל	0.2447	0.0004	0.0017	0.0035
איחל+שם פועל	0.0638	0.0423	-0.0181	-0.0800
איחל+פסוקית	0.1170	1.4207	0.1127	0.3630 (**TP**)

טבלה 9.17 : הפועל 'איחל'

Table 9.17 :'*aicl*'

verb-frame	RF	LLR	*t*-score	PMI
אמר+אל	0.0014	257.1612	-0.0586	-1.5709
אמר+את	0.0601	12259.7165 (**TP**)	-0.4298	-1.5197
אמר+ב	0.1593	1937.5997 (**FP**)	-0.1859	-0.4488
אמר+ל	0.0838	1307.9006 (**TP**)	-0.1497	-0.5312
אמר+מ	0.0100	1985.5841	-0.1636	-1.6068
אמר+על	0.0182	1544.2089 (**TP**)	-0.1505	-1.1516
אמר+עם	0.0067	293.7557	-0.0672	-0.8827
אמר	0.3302 (**FP**)	1536.8490 (**FP**)	0.1794	0.3034
אמר+שם פועל	0.0192	2148.8425 (**FP**)	-0.1756	-1.2835
אמר+פסוקית	0.4704 (**TP**)	44519.5152 (**TP**)	1.3080 (**TP**)	1.7542 (**TP**)

טבלה 9.18 : הפועל 'אמר'

Table 9.18 : '*amr*'

רשימת מקורות

Adler, M. (2007). *Hebrew Morphological Disambiguation:An Unsupervised Stochastic Word-based Approach*. (PhD thesis, Ben Gurion University of the Negev).

Atterer, M., & Schütze, H. (2007). Prepositional phrase attachment without oracles. *Computational Linguistics, 33*(4), 469-476.

Attia, M., Pecina, P., Tounsi, L., Toral, A., & van Genabith, J. (2011). Lexical profiling for arabic. *Proceedings of eLex,* , 23-33.

Baldewein, U., & Keller, F. (2004). Modeling attachment decisions with a probabilistic parser: The case of head final structures. *Proceedings of the 26th Annual Conference of the Cognitive Science Society,* 73-78.

Belletti, A., & Shlonsky, U. (1995). The order of verbal complements: A comparative study. *Natural Language & Linguistic Theory, 13*(3), 489-526.

Berman Aronson, R. (1978). *Modern hebrew structure.* Tel Aviv: Univ. Publ. Projects.

Brent, M. R. (1993). From grammar to lexicon: Unsupervised learning of lexical syntax. *Computational Linguistics, 19*(2), 243-262.

Briscoe, T., & Carroll, J. (1997). Automatic extraction of subcategorization from corpora. *Proceedings of the Fifth Conference on Applied Natural Language Processing,* 356-363.

Carroll, J. (1998). *Automatoc acquisition of subcategorization frames and selectional preferences from corpora.* Unpublished manuscript.

Carroll, J., Minnen, G., & Briscoe, T. (1998). Can subcategorisation probabilities help a statistical parser? *CoRR, cmp-lg/9806013*

Chesley, P., & Salmon-alt, S. (2006). Automatic extraction of subcategorization frames for french. *In Proceedings of the Language Resources and Evaluation Conference, LREC 2006,*

Chomsky, N. (1965). *Aspects of the theory of syntax*. Cambridge, MA: MIT Press.

Church, K. W., & Hanks, P. (1990). Word association norms, mutual information, and lexicography. *Computational Linguistics, 16*(1), 22-29.

Church, K., Gale, W., Hanks, P., & Hindle, D. (1991). Using statistics in lexical analysis. In U. Zernik (Ed.), *Lexical acquisition: Exploiting on-line resources to build A lexicon* (pp. 115-164). Hillsdale, NJ: Erlbaum.

Dahlgren, K., & McDowell, J. P. (1986). Using commonsense knowledge to disambiguate prepositional phrase modifiers. *AAAI,* 589-593.

Dunning, T. (1993). Accurate methods for the statistics of surprise and coincidence. *COMPUTATIONAL LINGUISTICS, 19*(1), 61-74.

Fisher, R. A. (1925). *Statistical methods for research workers*. Edinburgh: Oliver & Boyd.

Foth, K. A., & Menzel, W. (2006). The benefit of stochastic PP attachment to a rule-based parser. *Proceedings of the COLING/ACL on Main Conference Poster Sessions,* 223-230.

Garnsey, S. M., Pearlmutter, N. J., Myers, E., & Lotocky, M. A. (1997). The contributions of verb bias and plausibility to the comprehension of temporarily ambiguous sentences. *Journal of Memory and Language, 37*(1), 58-93. doi:10.1006/jmla.1997.2512

Goldberg, Y. (2011). *Automatic syntactic processing of modern hebrew.* (PhD thesis, Ben Gurion University of the Negev).

Hajič, J., Čmejrek, M., Dorr, B., Eisner, Y. D. a. J., Gildea, D., Koo, T., . . . Rambow, O. (2004). *Natural language generation in the context of machine translation*. Baltimore: Center for Language and Speech Processing, Johns Hopkins University.

Hindle, D., & Rooth, M. (1993). Structural ambiguity and lexical relations. *Computational Linguistics, 19*(1), 103-120.

Hirst, G. (1988). Semantic interpretation and ambiguity. *Artificial Intelligence, 34*(2), 131-177.

Ienco, D., Villata, S., & Bosco, C. (2008). Automatic extraction of subcategorization frames for italian. *Proceedings of the Sixth Language Resources and Evaluation (LREC'08),*

Itai, A., & Wintner, S. (2008). Language resources for hebrew. *Language Resources and Evaluation, 42*(1), 75-98.

Jensen, K., & Binot, J. L. (1987). Disambiguating prepositional phrase attachments by using on-line dictionary definitions. *Computational Linguistics, 13*(3-4), 251-260.

Korhonen, A. (2002). *Subcategorization acquisition.* ().University of Cambridge, Computer Laboratory.

Korhonen, A., Gorrell, G., & McCarthy, D. (2000). Statistical filtering and subcategorization frame acquisition. *Proceedings of the 2000 Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and very Large Corpora: Held in Conjunction with the 38th Annual Meeting of the Association for Computational Linguistics - Volume 13,* Hong Kong. 199-206. doi:http://dx.doi.org/10.3115/1117794.1117819

Korhonen, A., Krymolowski, Y., & Briscoe, T. (2006). A large subcategorization lexicon for natural language processing applications. *In Proceedings of LREC.*

Lapata, M., Keller, F., & Schulte im Walde, S. (2001). Verb frame frequency as a predictor of verb bias. *Journal of Psycholinguistic Reseach, 30(4),* 419-435.

Li, J., & Brew, C. (2005). Automatic extraction of subcategorization frames from spoken corpora. *Proceedings of the Interdisciplinary Workshop on the Identification and Representation of Verb Features and Verb Classes,* 74-79.

Lin, D. (1998). Dependency-based evaluation of MINIPAR. *Workshop on the Evaluation of Parsing Systems,* 317-330.

Messiant, C., Poibeau, T., & Korhonen, A. (2008). LexSchem: A large subcategorization lexicon for french verbs. *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08).*

Pantel, P., & Lin, D. (2000). An unsupervised approach to prepositional phrase attachment using contextually similar words. *Proceedings of the 38th Annual Meeting on Association for Computational Linguistics,* 101-108.

Ratnaparkhi, A., Reynar, J., & Roukos, S. (1994). A maximum entropy model for prepositional phrase attachment. *Proceedings of the Workshop on Human Language Technology,* 250-255.

Resnik, P., & Hearst, M. (1993). Structural ambiguity and conceptual relations. *Proceedings of the Workshop on very Large Corpora: Academic and Industrial Perspectives,* 58-64.

Roland, D., & Jurafsky, D. (1998). How verb subcategorization frequencies are affected by corpus choice. *Proceedings of the 17th International Conference on Computational Linguistics-Volume 2,* 1122-1128.

Ross, J. R. (1967). *Constraints on variables in syntax.* (PhD thesis, Massachusetts Institute of Technology).

Sarkar, A., & Zeman, D. (2000). Automatic extraction of subcategorization frames for czech. *COLING,* 691-697.

Schulte im Walde, S., & Brew, C. (2002). Inducing german semantic verb classes from purely syntactic subcategorisation information. *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics,* 223-230.

Shilon, R., Fadida, H., & Wintner, S. (2012). Incorporating linguistic knowledge in statistical machine translation: Translating prepositions. *European Chapter of the Association for Computational Linguistics (EACL 2012),* Avignon, France.

Shilon, R., Habash, N., Lavie, A., & Wintner, S. (2012). Machine translation between hebrew and arabic. *Machine Translation, 26*(1-2), 177-195.

Shlonsky, U. (1997). *Clause structure and word order in hebrew and arabic: An essay in comparative semitic syntax* Oxford University Press, USA.

Sima'an, K., Itai, A., Winter, Y., Altman, A., & Nativ, N. (2001). Building a tree-bank of modern hebrew text. *Traitment Automatique Des Langues, 42*(2)

Simon, E., Serény, A., & Babarczy, A. (2010). Automatic acquisition of hungarian subcategorization frames. *LREC,* 7-12.

Stetina, J., & Nagao, M. (1997). Corpus based PP attachment ambiguity resolution with a semantic dictionary. *PROCEEDINGS OF THE FIFTH WORKSHOP ON VERY LARGE CORPORA,* 66-80.

Surdeanu, M., Harabagiu, S. M., Williams, J., & Aarseth, P. (2003). Using predicate-argument structures for information extraction. *ACL,* 8-15.

Volk, M. (2002). Combining unsupervised and supervised methods for PP attachment disambiguation. *Proceedings of the 19th International Conference on Computational Linguistics-Volume 1,* 1-7.

Wilks, Y., Huang, X., & Fass, D. (1985). Syntax, preference and right attachment. *IJCAI,* 779-784.

Yeh, A. S., & Vilain, M. B. (1998). Some properties of preposition and subordinate conjunction attachments. *Proceedings of the 17th International Conference on Computational Linguistics-Volume 2,* 1436-1442.

Zeman, D. (2002). Can subcategorization help a statistical dependency parser? *Proceedings of the 19th International Conference on Computational Linguistics-Volume 1*, 1-7.

אזר, מ. (תשל"ב). סימני הצרכה, מלות הדרכה והיחידות המילוניות של הפועל. *לשוננו* (ל"ו), 220-227, 282-286.

בורוכובסקי-בר אבא א. (2001). *הפועל, תחביר, משמעות ושימוש, עיון בעברית בת זמננו*. הוצאת הספרים של אוניברסיטת בן-גוריון בנגב.

רובינשטיין, א. (תשל"ט). ערכיות תחבירית וערכיות סמנטית, *לשוננו* (מ"ג), 3-19.

שטרן, נ. (1994). *מילון הפועל, ערכיות ותפוצה של פעלים בעברית החדשה*. הוצאת אוניברסיטת בר-אילן, רמת-גן.

The result of our study is a complete Hebrew verb lexicon, as exists in other languages such as VALEX for English (Korhonen et al., 2006) or LexScheme for French (Messiant et al., 2008). For each verb, our lexicon includes its subcategorization frames along with statistical information about the occurrence of the verb and its subcategorization frames. Our verb lexicon includes 3,393 verb types, and 20,829 pairings of verbs with their subcategorization frame(s).

We hope that our research results will be useful to the scientific community of Hebrew language research, and serve as a basis for fruitful and extensive study of the verb.

behavior in a small subset of the corpus which had been manually morphologically annotated (the Hebrew TreeBank), we develop a technique which focuses on the words appearing immediately after the verb, and the required information is learned from this region. Nonetheless, although complements are usually close to the verb, they can be distanced by adjuncts that aren't necessarily required by the verb like place and time. Statistical hypothesis testing is used to learn the correct connection between the verb and its frames. We use by four statistical measures: 1. log-likelihood ratio (LLR); 2. pointwise mutual information (PMI), 3. t-test; and 4. raw frequency resulting in four different verb lexicons, one for each statistical measure.

We evaluate these lexicons in two ways: The first is an internal evaluation which tests the results on a small set of 17 verbs against a manual annotation. Based on this evaluation, we conclude that PMI is the best statistical measure for this task (with recall of 69.39% and precision of 97.14%). The second evaluation is external, through showing an improvement in the tasks of PP-attachment (reducing the error rate by 28.85%) and of machine translation from Arabic to Hebrew (increasing the BLEU score from 32.5 to 37, and the METEOR score from 52.6 to 56).

Technique #2: The second technique for extracting subcategorization frames uses a corpus which had been syntactically parsed using Goldberg's (2011) parser. We apply similar methods as in the first technique to extract subcategorization frames from this corpus, except instead of limiting our scope to the region immediately following the verb, (resulting in a single potential frame), we consider a larger set of partial subcategorization frames that are connected to the verb according to the syntactical parser.

The results of this technique are not good as those of the first technique (which used the morphologically-tagged corpus), perhaps because of the increase in the number of frames that could be connected to each verb. However, we believe that this second technique gives a more complete picture of the verb and its subcategorization frames, and could be improved with some small adjustments.

The linguistic phenomenon of subcategorization frames is quite complex. A subcategorization frame can be empty, with no complements required by the verb, as in the verb *mt* (מת, "die"). Alternatively, a frame can instead have several possible complements; for example, the verb *ht'rb* (התערב, "bet") has three prepositions that complete it, several combinations of which are acceptable: 1. *'m* (עם, "with"); 2. *'l* (על, "on"); and 3. *š* (ש, "that") (which corresponds naturally to the English "bet with/on/that"). There are verbs with more than one frame option, which can affect the verb's meaning. For example, the verb *qina* (קינא, "envy/jealous"), when paired with the preposition *l* (ל, "to") means "suspect infidelity of [someone]", but when paired with *b* (ב, "in") means "envy [something/someone]".

The relatively free word order in Hebrew presents a particular challenge, since verb complements may occur both before and after the verb, and is not always adjacent to the verb. Thus, subcategorization frames with several complements may appear in many distinct ways, making automatic identification more difficult.

To simplify the task, and because of the complexity of Hebrew and especially its free word order, we focus only on *partial subcategorization frames*, i.e., we restrict our attention to a single complement. In addition to limiting the size of the frame, we further constrain the problem by considering only common complements:

1. noun phrases;
2. prepositional phrases with one of the prepositions *'l* (על, "on"), *b* (ב, "in"), *l* (ל, "to"), *m* (מ, "from"), *'l* (על, "on"), or *'m* (עם, "with"), and not any other proposition;
3. complementizers with *š* (ש, "that") or *ki* (כי, "that"); and
4. infinitival verb phrases.

In our research we examine two techniques for recognizing the subcategorization frames:

Technique #1: The first technique uses a morphologically-tagged corpus. In such a corpus all we have is the morphological information of each word, and we need to discover the complements of the verb. Through studying the verb and its complements'

Abstract

The core of syntactic structure, according to most contemporary syntactic theories, revolves around verbs and their complements. Several linguistic theories map verb-complement constructions to semantic relations, encoding predicate-argument structures. Correctly identifying verb complements in naturally-occurring texts is therefore important both theoretically, for linguistic investigations, and practically, for natural language processing (NLP) applications. A crucial resource needed for this task is a dictionary of *verb subcategorization frames*, listing the number and types of complements that are most likely to occur with each verb, ideally with some statistical measure of the strength of the relation that holds between the verb and each of its complements.

Much research has been done on the automatic extraction of verb subcategorization frames in several languages, such as English (Korhonen et al., 2006) and French (Messiant et al., 2008), and even for Arabic (Attia et al., 2011), another Semitic language. To the best of our knowledge, Hebrew[1] has not yet been studied and this thesis is the first work which deals with automatic extraction of verb subcategorization in Hebrew.

Knowing the subcategorization frame of the verb is crucial for many applications, especially syntactic parsing. Studies have shown that about 50% of syntactic parsing errors result from lack of information about verb subcategorization frames (Carroll, 1998). Indeed, providing this information to a syntactic parser significantly improved its accuracy (Carroll, Minnen & Briscoe, 1998; Zeman, 2002).

[1]To facilitate readability, we use a Roman character transliteration of Hebrew: *abgdhwzxTiklmnš'pcqrst* (in Hebrew lexicographic order).

List of Figures

List of Tables

Contents

The Research Thesis Was Done Under the Supervision of Prof. Alon Itai and Prof. Shuly Wintner in the Department of Computer Science

The Generous Financial Help of the Technion is Gratefully Acknowledged

Automatic Extraction of Subcategorization Frames for Hebrew

Research Thesis

Submitted in partial fulfillment of the requirements

for the degree of Master of Science in Computer Science

Hanna Fadida

Submitted to the Senate of

the Technion - Israel Institute of Technology

Sivan 5772 Haifa June 2012

Automatic Extraction of Subcategorization Frames for Hebrew

Hanna Fadida