הערות למשתמש, גרסה 1.0

מוצר: מנתח מורפולוגי

ועדת היגוי ומחקר: אלון איתי, שולי וינטנר, עדו דגן, יועד וינטר, מיכאל אלחדד

פיתוח התוכנה: דליה בוז'ן

תיעוד: נעם אורדן

תאריך: 30 בנובמבר 2006

1. הקדמה 3
2. הניתוח 4
3. החוקים 5
3.1. חוקי נטיות 5
3.2. חוקי גזירה 5
4. מה בלקסיקון? 6
5. מה ב-*HebreWords* 6
6. קבצים מצורפים 7
7. הרצת המנתח 7
8. הוראות שימוש 8
8.1. מנתחים 8
8.1.1. מנתח טקסט (Text Analyzer) 8
8.1.2. מנתח XML (XML Analyzer) 8
8.2. מידע בנוגע לפרמטרים 8
9. על כמה החלטות שנתקבלו בנוגע ללקסיקון ולמנתח 9
בין תחביר למורפולוגיה 9
9.1. הבינוני 9
9.2. חיבור מילות זיקה 9
9.3. "עשרות": שם-עצם או כמת? 9
ענייני פעלים 10
9.4. ציווי בנפעל 10
9.5. צורת נתפעל 10
9.6. שתי צורות העבר (תבנית 61,62) 11
9.7. גזרת ע"ו 11
צורות ארכאיות 11
9.8. כינויי מושא חבורים 11
9.9. ה' השאלה 11
9.10. האם צריך לחולל את הצורות הבאות – אאמין ← אאמינה, אפקיד ← אפקידה? 12
שונות 12
9.11. הומונימים ופוליסמיה 12
9.12. תחיליות ופעלים 12
9.13. קניין של צורת "בנאי" 13
9.14. מספרים בגוף ראשון 13
9.15. קישורים לקסיקליים 13

1. הקדמה

המנתח המורפולוגי (*HAMSA*) פותח במרכז מילה (**מרכז ידע לתקשוב בשפה העברית**). המנתח מקבל כקלט תַּמְנִית (token) בעברית, ובתגובה נותן את כל אפשרויות הקריאה המורפולוגית של אותה תמנית. את הפלט ניתן לקבל או בצורת טקסט או בפורמט XML לפי סכימה הנגישה באתר האינטרנט של המרכז. (פרטים נוספים ניתן למצוא בכתובת http://yeda.cs.technion.ac.il:8088/mila/intro/index.shtml.)

```
מילה[+noun][+id]3265[+undotted]מילה[+transliterated]milh[+gender]+fem
inine[+number]+singular[+definiteness]+false[+register]+formal[+const
ruct]+false
מילה[+noun][+id]9414[+undotted]מילה[+transliterated]milh[+gender]+fem
inine[+number]+singular[+definiteness]+false[+register]+formal[+const
ruct]+false
מילה[+noun][+id]20302[+undotted]מיל[+transliterated]mil[+gender]+masc
uline[+number]+singular[+register]+formal[+construct]+false[+possessi
veSuffix]+3p/F/Sg
```

איור 1: פלט טקסט של המנתח: קריאות אפשריות של התמנית "מילה" (ניתן לקבל את הפלט גם בפורמט של XML

המנתח מטפל בצורות לא מנוקדות של עברית מודרנית. איננו מתיימרים לנתח כל טקסט עברי (למשל פרקי תנ"ך), ואיננו תומכים, בעיקרון, בצורות תחביריות שאינן קיימות בעברית מודרנית (כמו ו' ההיפוך). כך גם איננו מתחייבים לכסות ערכים לקסיקליים ארכאים שלא נפוצים בעברית מודרנית (אם כי ייתכן שישנן צורות כאלה שינותחו בהצלחה).

כרגע אנחנו תומכים באופן חלקי בכתיב לא תקני. לצד הצורה "בעיה", למשל, הוספנו גם "בעייה" ותייגנו אותה כבלתי-תקנית. איננו מטפלים בשגיאות כתיב. בשלב זה אנו מטפלים בכל מילה בנפרד, ועל כן, מילים המורכבות ממספר תמניות (MWT - MultiWordTokens) (כמו "בית ספר", או "שנים עשר") אינן זוכות לטיפול כמצרף.

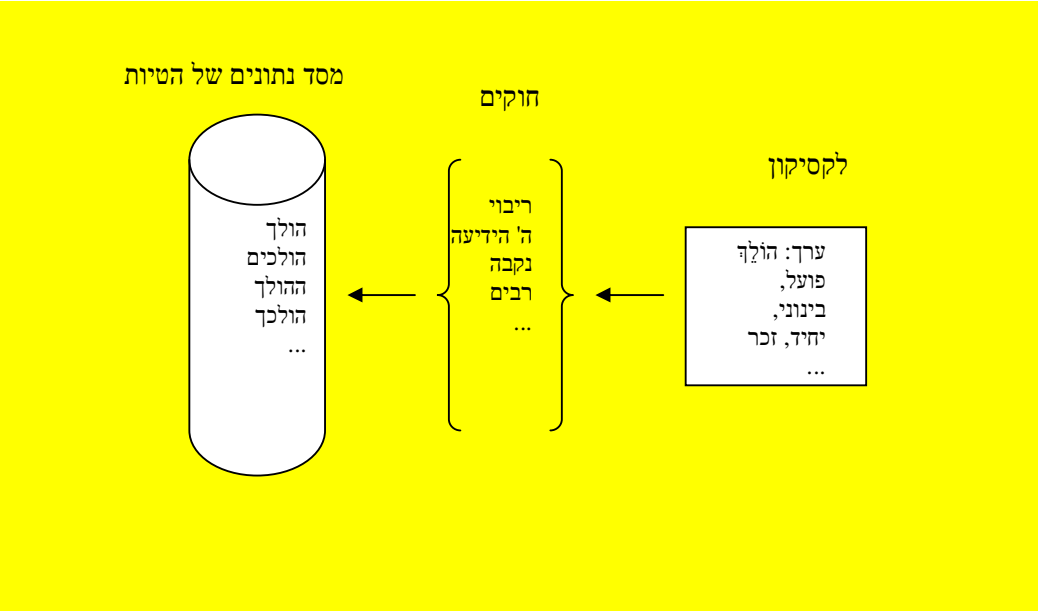
הניתוח מתבצע באמצעות שלושת הרכיבים של המנתח:

(1) לקסיקון אוסף של ערכים לקסיקליים בעברית, הכוללים מידע מורפולוגי מקיף (שחלקו יפורט להלן) על כל אחד מהערכים, ולעיתים אף מידע סמנטי, כמו הגדרה (שדה משמעות) או מקבילות תרגומיות באנגלית.

(2) חילול (generation) – המחולל מקבל ערך בלקסיקון, הכולל את הצורה ואת כל התכוניות (features) המורפולוגיות שלה (חלק דיבר, גוף, מספר וכו'), ובתגובה מייצר את כל ההטיות האפשריות של הערך. החילול נעשה על-פי אוסף של חוקים המפרטים כיצד ניתן להטות כל אחד מהערכים. כך, למשל, אם הצורה הבסיסית של הפועל מיוצגת בלקסיקון בתכוניות של גוף יחיד/עבר/נסתר, הרי שהחוקים מפרטים כיצד יש לייצר את כל הצורות

הרלבנטיות: עבור כל הגופים, כל הזמנים, כל כינויי הגוף, כמו גם צורות המקור, הבינוני וכן הלאה.

(3) מסד נתונים (*HebreWords Database*) המסד כולל למעשה את כל הצורות המוטות, היינו את כל ערכי הלקסיקון (כפי שמופיעים ב-1) ונטיותיהם (כלומר, לאחר שהוחלו עליהם החוקים של 2). מסד זה אינו כולל את אותיות השימוש (מש"ה וכל"ב).



איור 2: שלושת רכיבי המנתח

2. הניתוח

התחיליות החוקיות: העברית מאפשרת צירופים חוקיים שונים של מש"ה וכל"ב בהתאם לכל חלק דיבר. המנתח מכיר צירופים אלה ומפריד בינם ובין מחרוזת אותיות התואמת ערכים בלקסיקון.

אלגוריתם הניתוח:

1. עבור כל תמנית, מורידים כל תחילית חוקית של התמנית (כולל התחילית הריקה), ובודקים אם היא נמצאת בצורתה זו (ללא התחילית) ב-HebreWords;

2. אם היא נמצאת, וחלק הדיבר של המילה שנמצאה תואם את התחילית שהוסרה, הרי שיש עוד "קריאה" אפשרית. כך, למשל, הצורה "שבר" (לא מנוקדת) מופיעה כערך בלקסיקון (שֶׁבֶר) או כפועל (שָׁבַר), אך לאחר שמסירים את התחילית ש', מוצאים כי הערך "בר" (לשתיית אלכוהול) קיים כערך בלקסיקון, וכך יש קריאה נוספת ל"שבר".

3. החוקים

3.1. חוקי נטיות

באיור מספר 2 הוצגו שלושת רכיבי המנתח: הלקסיקון, החוקים ומסד הנתונים של ההטיות (HebreWords). הרכיב האמצעי של המנתח הוא אוסף של חוקים מורפולוגיים. כרגע ישנם 60 חוקים. כל חוק מתייחס לסוג מסוים של פריטים בלקסיקון ומותאם לתכונותיהם המורפולוגיות (חלק דיבר, מין וכו'); בהתאם, החוק מחולל את כל הנטיות האפשריות עבור כל ערך. יש לציין כי מדובר בנטיות אפשריות, אם כי לעיתים הם לאו דווקא סבירות. למשל, הכלל המתייחס לשייכות יחולל עבור שם העצם "ישיבה" את הצורה "ישיבתי", אם כי סבירות היקרותה בקורפוס נמוכה, ואולי אף לא קיימת (בניגוד ל"ראשי", למשל).

להלן שני חוקים פשוטים:

חוק מספר 5: יצירת ריבוי עבור שמות עצם המסתיימים באות ה': הסר את האות האחרונה והוסף את האותיות ות, למשל ארוחה ← ארוחות.

חוק מספר 17: יצירת סמיכות מצורת ריבוי המסתיימת במ"ם סופית: הסר את האות האחרונה, למשל אנשים ← אנשי.

כאמור, החלת כל החוקים על כל הערכים בלקסיקון מייצרת את HebreWords.

3.2. חוקי גזירה

נושא שלא נידון כאן כחלק מהמודל, אך ראוי להצגה מִקדמית, הוא הניסיון שלנו להעשיר את הלקסיקון באופן אוטומטי באמצעות חוקי גזירה. מדובר באוסף של חוקים (כרגע ישנם כ-15 חוקים שכאלה), שבאמצעותם אנו לוקחים ערכים קיימים בלקסיקון, מפעילים עליהם חוקי גזירה, וכך יוצרים ערכים חדשים. דוגמה: על פעלי בניין התפעל אנו מחילים את החוק הפשוט הבא כדי ליצור את שם הפעולה שלו:

1. הצורה הבסיסית: צורת העבר + וּת, למשל: הִתְפַּעֵל + וּת ← הִתְפַּעֲלוּת. ע' הפועל מקבלת דגש חזק ("דגש תבניתי"), אלא אם כן היא גרונית (א', ה', ח', ע', ר'). בנוסף
   1.1. הניקוד ב-ע' הפועל משתנה לשְׁוָא (ְ): הִשְׁתַּמֵּר ← הִשְׁתַּמְּרוּת.
   1.2. אם ע' הפועל היא א', ה', ח', או ע' (כלומר, גרוניות בלי ר'), היא נחטפת בשם הפעולה: הִתְפַּעֲלוּת, הִתְפָּאֲרוּת וכו'.
2. אלא אם כן
   2.1. ל' הפועל ה', שאז היא נשמטת ← הִתְבַּלּוּת, אלא אם כן
      2.1.1. ל' הפועל היא ה' שנהגית, שאז יש לשנות בהחלפה ← הִתְגַּבְּהוּת.

החוקים האלה אינם חלק מהמוצר שלנו, אלא מתודולוגיית עבודה.

4. מה בלקסיקון?

הלקסיקון כולל 22179 פריטים, וטבלת ההטיות כוללת של פריטי הלקסיקון כוללת 7336645 עיולים (entries).

אלה חלקי הדיבר המכוסים על ידי הלקסיקון:

חלק דיבר	מספר עיולים בלקסיקון
שמות עצם (nouns)	11603
פעלים (verbs)	4422
שמות פרטיים (proper names)	3253
שמות תואר (adjectives)	2311
תוארי פועל (adverbs)	403
מילות יחס (prepositions)	91
כינויי גוף (pronouns)	78
מילות חיבור (conjunctions)	69
מספרים (numerals)	62
מילות קריאה (interjections)	41
כמתים (quantifiers)	33
מילות שאלה (interrogatives)	9
מילות שלילה (negations)	5

5. מה ב-*HebreWords*

ההטיות הבאות מכוסות ב-*HebreWords*:

- שם-עצם: יחד/רבים, נפרד/נסמך, שייכות; אם רלבנטי, מציינים גם את צורת הנקבה ואת הזוגי.
- תואר-השם: יחד/רבים, זכר/נקבה, מקור, שייכות (אם יש, כמו "יפתי").

- פועל: פועל נטוי (infinite verbs), גוף, מספר, זמן (עבר, עתיד, ציווי, שם פעולה), הצורה השמנית של הפועל (infinite verbs): בנטייה למושא או נושא. הפעלים מופרדים לפי בניינים.
- תואר הפועל: אין נטיות, חוץ מיוצאי דופן נדירים (אותם אנו מציינים), כמו "לאיטו", "לאיטך" וכו'.
- שם פעולה, כולל נטייה לשייכות.
- מילת יחס: יחיד/רבים, זכר/נקבה, גוף.
- כמתים: לעיתים נוטים לזכר/נקבה ולגוף. לעיתים נדירות נוטים למספר (אנו מציינים זאת במקרים אלה).
- מספרים: סודרים ומונים, כולל הנטיות הרלבנטיות. למספרים מונים: זכר/נקבה, נפרד/נסמך. למספרים סודרים: זכר/נקבה ('ראשון'/'ראשונה' נוטים לנפרד/נסמך). שברים: נוטים לנפרד/נסמך ולשייכות.

6. קבצים מצורפים

הגרסה הנוכחית כוללת את הקבצים הבאים:

- קובץ הטיות – טבלה אחת ענקית הנמצאת ב -HebreWords.
- קובץ תחיליות – טבלה נפרדת הכוללת את כל הצירופים החוקיים של מש"ה וכל"ב.
- קובץ גימטריה - טבלה נפרדת הכוללת את כל המספרים 1-999 מיוצגים באותיות עבריות.

7. הרצת המנתח

את המנתח ניתן להריץ בשני אופנים:

1) שליפת ההטיות ממסד נתונים, מטבלאות המכילות את כל ההטיות. זו דרך העבודה שלנו כאשר מישהו מגיש שאילתת ניתוח דרך האינטרנט.

2) שליפת ההטיות מקבצי data. הקבצים אינם מצריכים התקנת MySQL או כל כלי אחר המטפל במסד נתונים. זה הכלי שאנו מעמידים לרשות המשתמשים שלנו.

8. הוראות שימוש

8.1. מנתחים

הוראות הרצה למנתח תמניות (tokenizer):

הרצת מנתח התמניות היא הכרחית על מנת להריץ את המנתח המורפולוגי. המנתח יודע לטפל רק בטקסטים שעברו טוקניזציה.

המנתח המורפולוגי ומנתח התמניות נבדקו בסביבת "חלונות" ובסביבת לינוקס: אין הבדל באופן ההרצה. כמו כן, ניתן לעבוד ניתן גם לעבוד ב-pipeline.

8.1.1. מנתח טקסט (Text Analyzer)

java -Xmx1024m -jar TextAnalyzer.jar FALSE

[inputFile.txt][outputFile.txt][dinflectionFile][PrefixesFile][GimatriaFile]

8.1.2. מנתח XML (XML Analyzer)

java -Xmx1024m -jar XMLAnalyzer.jar FALSE

[inputFile.txt][outputFile.txt][dinflectionFile][PrefixesFile][GimatriaFile]

8.2. מידע בנוגע לפרמטרים

- False מתייחס לאפשרות הגישה למסד הנתונים (true/false), יש להשתמש אך ורק ב-false. יש צורך בפרמטר זה מאחר שיש שני אופני הרצה – מול מסד נתונים ומול קבצי data (ר' הסבר בהקדמה). בשלב זה כל המשתמשים שלנו עובדים מול קובצי data, ולכן עובדים אך ורק ב-false.
- [inputFile] נוגע לנתיב המלא של קובץ הקלט. יש לקדד את קובץ הקלט ב-UTF8.
- [inflectionsFile] מספק את הנתיב המלא לקובץ inflections.data.
- [outputFile] נוגע לנתיב המלא של קובץ הפלט; עבור XML הסיומת תהיה *.xml.
- [prefixesFile] מספק את הנתיב המלא של קובץ prefixes.data.
- [gimatriaFile] מספק את הנתיב המלא של הקובץ gimatria.data.

9. על כמה החלטות שנתקבלו בנוגע ללקסיקון ולמנתח

בין תחביר למורפולוגיה

9.1. הבינוני

לצורת הבינוני תפקידים תחביריים שונים: פועל בהווה, שם-עצם או שם-תואר. ההחלטה איזה תפקיד יש למילה נקבעת רק בניתוח תחבירי של המשפט כולו, ואי אפשר להסיקו רק על פי המילה עצמה. על-כן, ניתחנו מילים אלה כבינוני והותרנו למנתח התחבירי להבחין בין התפקידים התחביריים השונים.

דוגמא:

שובר[+participle][+id]20136[+undotted]שבר[+transliterated]ebr[+root]שבר[+binyan]+
Pa'al[+register]+formal[+tense]+beinoni[+person]+any[+gender]+masculine[+numbe
r]+singular[+construct]+false[+definiteness]+false

לפעלים יוצאים של בניין קל קיימת גם צורת הבינוני הפעול, המנותחת כדלקמן:

שבור[+participle][+id]20136[+undotted]שבר[+transliterated]ebr[+root]שבר[+binyan]+
Pa'al[+register]+formal[+tense]+beinoni[+person]+any[+gender]+masculine[+numbe
r]+singular[+construct]+false[+definiteness]+false

9.2. חיבור מילות זיקה

ברמת התחביר יש להבדיל בין ה' הידיעה (definite article) לבין ה' הזיקה (relativizer).

- נהג <u>ה</u>נתפס במהירות מופרזת <u>וה</u>מבקש להמיר את הקנס במשפט יוכל לפנות לבית המשפט תוך שלושים יום.

המנתח אינו מבחין ביניהם, כי נראה שהבחנה היא תחבירית (ולא מורפולוגית), ולכן ראוי להשאיר את הטיפול בכך למנתח תחבירי.

9.3. "עשרות": שם-עצם או כמת?

לכאורה, לפחות, נראה כאילו עשרות מתנהג גם כשם עצם. בניגוד לכמתים כמו "כמה", ניתן לראות ב"עשרות" ש"ע הנוטה לקניין.

1.א. כמה אנשים הגיעו מתאילנד.

1.ב. *כמה של אנשים הגיעו מתאילנד.

לעומת:

2.א. עשרות אנשים הגיעו מתאילנד.

2.ב. עשרות של אנשים הגיעו מתאילנד.

כפי שניתן לראות מהניגוד: (1ב) אינו דקדוקי ואילו (2ב) דקדוקי וקביל (ה"מבחן" הזה נעשה לפי Glinert). מאידך, מבחינה תחבירית "עשרות" אינו מתנהג כ-head, מאחר שאנו אומרים "עשרות אנשים מגיעים", ולא "עשרות אנשים מגיעות". לפיכך הוחלט לתייג "עשרות", "מאות", "אלפי" וכו' ככמתים, ולהשאיר את "ההתמודדות התחבירית" למנתחים תחביריים.

ענייני פעלים

9.4. ציווי בנפעל

מובן שחלק מהמשמעויות של נפעל הן "פעילות" (נכנס) ולכן ניתן לצוות: "היכנס!". אך נראה שגם על חלק מהמשמעויות הסבילות ניתן לצוות: "החנק!" או "הירגע!" או "היתלה!". ניתן לשקול להוסיף בעתיד היוריסטיקות: למשל, בשורשים המתממשים גם בפעל וגם בנפעל, סביר להניח שהמשמעות של נפעל סבילה, ולכן אין סבירות גבוהה שלציווי תהיה משמעות.

ובכל זאת, בשלב זה המנתח מייצר ציווי לכל הפעלים בבניין נפעל.

9.5. צורת נתפעל

בעיקרון צורת נתפעל היא שריד של בניין שכבר לא קיים (ראה ביאליק "החמה נסתלקה"). יחד עם זאת, לעיתים יש בידול משמעות עדין בין צורה זו ובין התפעל, כמו במשפט "נזדמנתי למקום". בכל מקרה, יצירה של צורה זו יכולה אף היא להוביל לחילול-יתר, באשר "נתפעל" יכול להיות גם גוף-שלישי/יחיד/עבר וגם גוף-ראשון/רבים/עתיד. לכן הוחלט, בשלב זה, להוסיף צורות אלה באופן ידני בתאם למימושם בקורפוס. בשלב זה יש 485 בצורת נתפעל.

9.6. שתי צורות העבר (תבנית 61,62)

עבור חלק מהפעלים הן מתקיימות, כמו "נבוכתי" ו"נבוכותי" (למרות שרוב הדוברים העבריים יאמרו היום "הובכתי" בהופעל), או "נסוגתי" לצד "נסוגותי" (חיפוש באינטרנט מאשש הנחה זו). במקרים אלה אין הבדל במשמעות.

9.7. גזרת ע"ו

בבניינים הכבדים קורה שאותו פועל מתממש בשתי צורות שונות בגזרת ע"ו, כמו, למשל, "התעורר" ו"התעוור" (שניהם בעלי שורש ע.ו.ר.). הלקסיקון משקף כל צורה בעיול נפרד. עיולים אלה הוזנו באופן ידני.

צורות ארכאיות

9.8. כינויי מושא חבורים

מדובר על מקרים כמו :

- אחתכנו – אחתוך אותו
- ארפאך – ארפה אותך

החלטנו שלא לכלול כינויי מושא חבורים של פעלים סופיים במנתח. לפי בדיקה שערכנו, מתברר שהצורות האלה נדירות מאד בעברית מודרנית, ורובן, כאשר מתממשות, מתממשות במסגרת ביטויים קפואים, שאותם ניתן להוסיף באופן ידני. הוספה של מבנה זה תגדיל שלא לצורך את טבלאות המנתח, ותוסיף רב משמעות רבה, למשל, האם "אכלנו" הוא גוף ראשון רבים או "[הוא] אכל אותנו"? עם זאת, הכללנו את כינויי המושא (והנושא) החבור לשם הפועל. ("שֶבתוֹ", "להשיגך" וכן הלאה) משום שהם יותר נפוצים.

9.9. ה' השאלה

גם ה' השאלה לא נכללה. על אף האופנה האחרונה להציע נישואים בשאלה "התינשאי ליז?", מצאנו כי ה' השאלה אינה נפוצה בעברית מודרנית ויוצרת יותר מדי רב-משמעות, ולכן המנתח שלנו אינו מציע אותה בניתוח.

9.10. האם צריך לחולל את הצורות הבאות – אאמין ← אאמינה, אפקיד ← אפקידה?

בעיקרון מדובר הצורות ארכאיות, נדירות למדי בעברית מודרנית. אאמינה, למשל, מניב 71 לעומת 17800 תוצאות בגוגל). הוחלט להוסיף צורות אלה רק לפעלים שבהם יש ראיות מובהקות להתממשות הצורה בקורפוס, כמו "הבה **נרעישה** ברעשנים".

שונות

9.11. הומונימים ופוליסמיה

נהוג להבחין בין מילים הומונומיות ומילים פוליסמיות:

- פוליסמיה (רב-משמעות) – בין המילים מתקיים דימיון מורפולוגי, אך הם נבדלים במשמעותן, כאשר נראה כי משמעות אחת נגזרת מהשנייה, כמו "הלך": 1. התקדם ברגל; 2. חלף ונעלם ("הגשם חלף הלך לו").
- מילים שיש להן צורה זהה, אך נראה כי משמעות אחת לא נגזרה מהשנייה, כמו "אשפה" (זבל) ו"אשפה" (של חיצים).

מאחר שהלקסיקון הוא לקסיקון מורפולוגי, כל עוד אין שום הבדל בנטייה, בשני המקרים (הן פוליסמיה הן הומונימיות) נתקין עיול אחד בלבד. יחד עם זאת, אנו מכלילים את שתי המשמעויות בשדה המשמעות, או שני תרגומים שונים, במידה שבאנגלית מדובר במקבילות תרגומיות שונות. כך, למשל "מקור" (של מים, source) ו"מקור" (של ציפור, beak) נוטים אחרת לרבים: הראשון – "מקורות", השני – "מקורים". בנוסף, למילה "מקורות" ברבים יש משמעות מובחנת (מקורות טקסטואליים יהודיים, כמו בביטוי "מצאנו במקורות"). מילה זו נכללת בלקסיקון, אך לא משום שיש לה משמעות שונה, אלא משום שבמשמעות זו היא מתממשת ברבים וברבים בלבד, ולכן "זכאית" לערך נפרד. בה-במידה חברת "מקורות" (proper name) זוכה אף היא לעיול נפרד.

9.12. תחיליות ופעלים

- באופן עקרוני, פעלים בציווי אינם מקבלים כל תחילית. יחד עם זאת, מצאנו שיש כמה יוצאי דופן: ו' החיבור ("הסכת ושמע"), או ב' "אדוורביאלית", כמו בביטוי "נשק בהכתף". כל עוד מדובר בביטויים הגדולים מתמנית אחת, יש לטפל בהם לחוד במסגרת MWT (מילים בעלות יותר מתמנית אחת).

- ה' הידיעה אינה מצטרפת לפעלים, אך היא יכולה להצטרף לבינוני.

9.13. קניין של צורת "בנאי"

לפי רב-מילים אין קניין ליחיד, אך לפי ספר השמות של ברקלי היו"ד נשמטת: "ספורטאך" (הספורטאי שלך). נראה כי יש לצדד בגישה של רב-מילים (ברקלי אף מציין בהקדמה שבספרו מופיעות "צורות תיאורטיות", היינו כאלה שלא מומשו אך אולי ימומשו בעתיד, ואילו המנתח שלנו עוסק רק בצורות קיימות).

9.14. מספרים בגוף ראשון

מספרים מונים מקבלים כינויים חבורים, כמו "שניכם", "ארבעתנו" וכן הלאה. אלה נוספו ידנית למספרים המונים. אולם כאשר הכינוי בא בגוף ראשון או שני, הוא תויג ככינוי גוף, מאחר שזהו תפקידו: "שנינו", "שלושתנו" כמו גם "שניכם", "שלושתכם" וכו'. אגב, רב מילים מנתח רק עד 6 ("שִׁשְׁתֵּנוּ"), ואכן צורה כמו "שבעתנו" היא ודאי נדירה ביותר (ואף יוצרת רב-משמעות, שכן ניתן לנתחה גם כ-שֶׁבְּעִתֵּנוּ).

9.15. קישורים לקסיקליים

מבחינה סמנטית המילים "שלישייה", "רביעייה" וכו' קשורות למספרים. מאידך, דקדוקית מדובר בשמות עצם לכל דבר. לכן הוסף שדה נוסף לשדות העיול של ש"ע: **קישור לקסיקלי**. בשדה זה ניתן לציין עבור הערך "שלישייה" את המספר המזהה של שם המספר הרלבנטי (במקרה הזה 23992 , "שלוש"). שדה זה יכול לשמש גם בקישורים לקסיקליים/סמנטיים אחרים, כמובן.

ספרות

Alon Itai, Shuly Wintner and Shlomo Yona. **A Computational Lexicon of Contemporary Hebrew**. In *Proceedings of LREC-2006*, Genoa, Italy, May 2006.

Shlomo Yona and Shuly Wintner. **A finite-state morphological grammar of Hebrew**. *Natural Language Engineering* 2007 (forthcoming).

Shlomo Yona. **A finite-state based morphological analyzer for Hebrew**. November 2004. University of Haifa M.Sc. Thesis.