כתב העת אוריינות ושפה
כרך 6 (2017) 22-36

שכיחות פונמות ורצפי פונמות בעברית בתהליך מונחה-נתונים

ורד זילבר-ורוד[2,1], משה לטין[3] ועמי מויאל[3]

[2]האוניברסיטה הפתוחה ; [3]אפקה - המכללה האקדמית להנדסה בתל-אביב

מאמר זה מתאר מחקר יישומי שנועד בראש ובראשונה להוות בסיס להקמת תשתית לשונית לאימון מודלים אקוסטיים ומודל לשוני עבור מנוע זיהוי דיבור (Automatic Speech Recognition (ASR) בשפה העברית. השלב המתואר במאמר מתייחס לתהליך מיצוי המשפטים שעתידים להיות מוקראים ומוקלטים על ידי דוברי עברית. בתהליך זה נאספו 50,000 משפטים מגוונים ודחוסים פונטית בגישה מונחית-נתונים (data driven approach), מתוך קורפוס טקסטואלי המכיל מעל 130 מיליון תמניות (כ-1 מיליון מילים תבניות). כל המשפטים עברו תיעתוק על פי לקסיקון מוכן מראש, כלומר נערכה המרה אוטומטית של הכתיב האורתוגרפי של המשפט לרצף הפונמות שמייצגות אותו ושהוחלט עליו מראש. ניתוח הממצאים כולל תמונה מייצגת של שכיחות 31 הפונמות בשפה העברית, שכיחות צמדי פונמות (diphones), ושכיחות רצפים של שלוש פונמות (triphones), אשר התקבלו ברמת המשפט, כלומר בין-מילים ובתוך-מילים. הממצאים מראים כי הפונמה השכיחה ביותר היא /a/ ואילו הנדירה ביותר היא /ʒ/. עוד נמצא כי העיצור [p] נמצא במיקום ה-27 ברמת השכיחות ואילו עיצורים מסוננים יותר, כמו [x] ו-[k] נמצאו במיקומים גבוהים יותר (22 ו-12, בהתאמה). הנתון על שכיחותו הנמוכה של העיצור [p] מפתיע גם לאור שכיחותו הגבוהה בשפות העולם. כמו כן נמצאו דיפונים וטריפונים מגוונים וחדשים, כלומר הקשרים מגוונים לכל הפונמות, מעבר לרמת ההברה והמילה.

מילות מפתח: שכיחות פונמות, רצפי פונמות, תהליך מונחה-נתונים, נתוני עתק, זיהוי דיבור, שפה עברית.

מבוא

מהי תפוצת הצלילים בעברית ומהם רצפי הצלילים הקיימים בה? על השאלות האלה מנסה המאמר הנוכחי לענות. תחום המחקר שעוסק בשאלות מסוג זה הוא הטיפולוגיה הפונולוגית, שמטרתו לבחון את השונות בין השפות ומנגד - את האוניברסלים הלשוניים. העיקרון האוניברסלי שמכתיב את תפוצת הפונמות בשפה מסוימת הוא "היררכיית המסוננות" (Schmid, 2012). בהקשר של מצאי הפונמות, מסוננות, שהיא מושג מפתח בטיפולוגיה בלשנית, משמעותה היא ששפה שמכילה צלילים מורכבים ומסוננים, תכיל בוודאות גם צלילים פחות מורכבים (כלומר, שיש בהם פחות תכוניות), שהם גם פחות מסוננים. לדוגמה, כל שפה שמכילה צליל יחסית מורכב כגון החוכך הווילוני [x], תכיל גם את הסותם הבלתי-קולי הפחות מסוננן [p]. על פי תאוריית הפונולוגיה כהתנהגות האדם (Phonology as Human Behavior – PHB), אנחנו צופים שהעיצורים הקלים ביותר להגייה יהיו גם הנפוצים ביותר (Wolf & Tobin, 2011). עקרון המסוננות עומד ביסודה של "היררכיית העדיפויות" במצאי הפונמות של כל שפה. האחרון הוא מושג שטבעו Maddieson (1984) ו-Maddieson and Precoda (1990) לאחר שחקרו כ-460 שפות. דוגמה לדירוג עדיפויות של מצאי פונמות היא שבכל שפה יש לפחות שתי תנועות [a] ו-[i]. אם בשפה יש שלוש תנועות,

[1] vereds@openu.ac.il

יהיו שתי אלה וגם [u]. משמעות הדבר היא שתפוצת הפונמות בשפה אינה שרירותית. Hay and Bauer (2007) אף הראו כי יש קשר בין מצאי הפונמות בשפה לבין גודל האוכלוסייה הדוברת אותה (את מחקרם הם עשו גם על העברית). אולם יש להדגיש כי הפן הפיזיולוגי של הפונמות (כלומר, קיומן של התכוניות החיתוכיות ומידת מורכבותן) אינו ההסבר היחיד למצאי הפונמות בשפה. אילו ההסבר היחיד היה פיזיולוגי-טבעי, כי אז היינו מצפים מכל הצלילים בכל השפות ל"התנהגות" דומה. לטענת Schmid (2012: 46-47) יחסי מסומננות נגזרים גם משכיחויות הפונמות, כלומר שבחינת שכיחויות של פונמות בשפה חושפת גם את יחסי המסומננות בין הפונמות.

מחקר יישומי כתשתית לטכנולוגיות עיבוד דיבור

מאמר זה מתאר מחקר יישומי, שנועד בראש ובראשונה להוות בסיס להקמת תשתית לשונית למנוע זיהוי דיבור (Automatic Speech Recognition, ASR) בשפה העברית.[2] מנוע זיהוי דיבור מזהה, בראש ובראשונה, את ההגאים הנאמרים, על פי מודלים אקוסטיים שהוא אומן להם והזיהוי האוטומטי מתבסס גם על מודל שפה (language model) ברמה הצלילית, שמתבסס על הסתברויות סטטיסטיות של רצפי צלילי הדיבור המופיעים בשפה (Lamel, Kassel, & Seneff, 1989). מודלים אקוסטיים של מימושי הפונמות ושל סביבתם (context), ברמה של הצליל הבודד (monophone), 2 צלילים (diphones. להלן: דיפון) עד 3 צלילים (triphones, להלן: טריפון), הינם תשתית חיונית לכל טכנולוגיית עיבוד דיבור, בראש ובראשונה טכנולוגיית זיהוי דיבור (Automatic Speech Recognition), הידועות גם בכינוי: "דיבור לטקסט" (Speech to Text (STT)), אך גם לטכנולוגיית הדיבור המלאכותי המכונה: "טקסט לדיבור" (Text to Speech (TTS)).

החשיבות בייצוג דיפונים וטריפונים בהקלטות לאימון הפונמות מקורה בעובדה שחוקי צירופי הפונמות (המכונים פונוטקטיקה. ראו לאופר (תשנ"א)) בתחום ההברה הם חוקים תלויי שפה. לכל שפה החוקים האופייניים לה ותפוצת הפונמות והאלופונים נובעת מחוקים אלה. חוקים פונוטקטיים, למשל, קובעים אילו פונמות יכולות להופיע בעמדת האונסט (onset, ראש ההברה) ואילו בעמדת הקודה (coda, סוף ההברה), אילו פונמות יכולות להופיע בסמיכות זו לזו, ואילו לא. סדר הפונמות בצרור העיצורים בעמדת האונסט והקודה נתון אף הוא לחוקים פונוטקטיים, חלקם אוניברסליים (כגון עקרון רצף הצליליות (SSP – Sonority Sequencing Principle)) וחלקם תלויי שפה.

על פי לאופר (תשנ"א) מבנה ההברה בעברית המודרנית הוא: (CCC)V(CCC), כלומר כל הברה חייבת להכיל לפחות תנועה (V), שהיא גרעין ההברה, ובמשבצות של האונסט והקודה יש מקום ל-0 עד 3 עיצורים. ואולם, צרורות העיצורים צריכים לציית לעקרון רצף הצליליות. עיקרון זה אומר כי "אין עלייה בסונוריות מגרעין ההברה אל קצותיה" (אדם, 2014, עמ' 143). תנועות הן הצליליות ביותר, מאפיין שנובע מכך שאין בהגייתן התקרבות או התחככות של איברי דיבור, כלומר מסלול הקול פתוח ולכן עוצמת הצליל המופק היא הגבוהה ביותר יחסית לעוצמה בעת הגיית צלילי דיבור אחרים. פחות צליליים מן התנועות הם הגאי מעבר (כגון [y], ואז, בסדר יורד ובחלוקה גסה לקבוצות: עיצורים שוטפים (כגון [l]), אפיים, חוככים, ולבסוף הסותמים. בלוח 1 מומחש העיקרון במילה הדו-הברתית 'פליטה' [plita]. החיצים כלפי מעלה

[2] מאמר זה הוא עדכון של זילבר-ורוד, לטין ומויאל (2013).

ממחישים את העלייה בצליליות. שני העיצורים הראשונים מהווים צרור עיצורים באונסט של ההברה הראשונה. העיצור הסותם /p/ פחות צלילי מהעיצור /l/, וגרעין ההברה /i/ הוא הצלילי ביותר. גם בהברה השנייה /ta/ העיקרון נשמר, כפי שממחיש החץ. בעברית, ובשפות אחרות, מתרחשות הפרות של העיקרון הזה, חלקן נפוצות מאוד, כמו התופעה של הופעת עיצור שורק (strident) לפני סותם (למשל, במילה 'סטירה' [stira]), אולם ההנחה היא כי עיקרון זה משפיע על חלוקת המילה להברות ומהווה טריגר להופעתן של תופעות פונולוגיות במבני הברה שונים.

לוח מס' 1
המחשה של עקרון רצף הצליליות במילה "פְּלִיטָה"

סולם הצליליות	p	l	i	.	t	a
תנועה			i			a
חצי תנועה						
שוטפים		l				
אפיים						
חוככים						
סותמים	p				t	

בכדי להגיע לייצוג של כל הפונמות וסביבתן חיוני שבסיס הנתונים הקולי יתבסס על כמות הופעות (occurrences) מספקת לכל פונמה וחזרות מרובות בצורה מספקת של כל הדיפונים והטריפונים שאופייניים לשפה (Gibbon, Moore, & Winski, 1997). תהליך האימון מצריך מספיק הופעות של כל צליל בבסיס הנתונים לאימון ומספיק חזרות של כל דיפון וטריפון שאופייניים לשפה, כדי למקסם את הסביבות הפונטיות ולהפיק גם את המימוש האלופוני של כל פונמה.

יש שתי דרכים שבהן גורמים להגייה של ייצוג כזה.

1. קורפוס משפטים "מאוזן פונטית" (phonetically balanced), עם ייצוג יחסי, לפי השכיחות בשפה. היתרון בשיטה זאת הוא בהסתמכות על משפטים שנשמעים יחסית טבעי ואפשרות להפיק את המשפטים מתוך טקסט קיים (דיבור או כתב); החיסרון: פונמות נדירות בשפה לא ייוצגו מספיק פעמים (למשל, /w/ בעברית).
2. קורפוס משפטים "עשיר פונטית" (phonetically rich), עם ייצוג כמעט זהה לכל הפונמות. משפטים עשירים פונטית נקראים בספרות המחקרית גם משפטים דחוסים פונטית (phonetically compact). היתרון בשיטה הוא שיש בקרה ושליטה מראש על כל התשתית הלשונית; החיסרון: החמצה של הגייה אלופונית של הפונמה בסביבות נוספות ברמת המשפט.

בפרויקט הדגל של אימון אקוסטי-פונטי בשפה האנגלית-אמריקאית, התמודדו עם החסרונות והיתרונות של שתי הדרכים באופן שיתואר להלן. פרויקט TIMIT – Acoustic-Phonetic Continuous Speech (Garofolo et al., 1993; TIMIT), הוא פרויקט של הקלטות שנעשו לצורך אימון 60 הפונמות באנגלית-אמריקאית בתחילת שנות התשעים של המאה ה-20. במסגרתו הוקלטו 6,300 משפטים על ידי 630 דוברים שמייצגים 8 דיאלקטים באנגלית-אמריקאית. כל דובר קרא כ-10 משפטים, כמתואר בלוח 2. ב-(1) מובא משפט לדוגמה:

She had your dark suit in greasy wash water all year. (1)

הפרויקט נועד למחקרים פונטיים-אקוסטיים של השפה האנגלית-אמריקאית, ולפיתוח והערכה (evaluation) של מנועי זיהוי דיבור. הפרויקט מהווה סימוכין לאימון פונטי בכל שפות העולם.

שיטת איסוף המשפטים כללה שלושה סוגים, שנועדו לספק איזון בין הדרישות הסותרות של: כיסוי דחיסות פונטית; הקשר רחב ומגוון; ושונות בין דוברים:

1. משפטי "שיבולת"[3] מכוילים (caliberated), שנועדו להציף הגיות של הדיאלקטים השונים.
2. משפטים דחוסים-פונטית (phonetically-compact)[4], שנועדו ליצור ייצוג של כל הדיפונים בשפה ועם ייצוג מופלג של סביבות פונטיות קשות או מעניינות (Garofolo et al., 1993). המשפטים נוצרו על ידי תכנון וניסוח מראש, תוך הקפדה על כיסוי מלא של דיפונים.
3. משפטים מגוונים-פונטית (phonetically-diverse), שנלקחו, כמו בשיטה של המחקר הנוכחי, מבסיס נתונים טקסטואלי, כדי לגוון את המשפטים, כדי לתת ייצוג להקשרים פונטיים נוספים וכדי לספק הקשרים חלופיים והיקרויות חוזרות של אותן סביבות פונטיות, אך במילים שונות.

כל דובר קרא את שני המשפטים הדיאלקטליים (כלומר, שנועדו להציף הבדלים דיאלקטליים באנגלית) וכל 630 הדוברים קראו אותם, כך שהוקלטו כ-1,260 מבעים. כל דובר קרא 5 משפטים דחוסים-פונטית, וכל משפט מ-450 הדחוסים-פונטית הוקרא על ידי שבעה דוברים, כך שהוקלטו כ-3,150 מבעים. כל דובר הקריא 3 משפטים מגוונים, שנבחרו אקראית מקורפוס טקסטואלי, וכל משפט מ-1890 המשפטים המגוונים הוקרא רק פעם אחת, כך שהוקלטו 1890 מבעים. סך הכול, כל אחד מ-630 הדוברים שגויסו להקלטות קרא כ-10 משפטים. לוח 2 מתאר את הנתונים על פי כמות המשפטים וכמות הדוברים שהגו כל משפט, וסה"כ המבעים שנהגו בפרויקט כולו (Zue, Seneff, & Glass, 1990).

לוח מס' 2

סוגי המשפטים ותכנית ההקלטות בפרויקט TIMIT

סוג משפט	כמות המשפטים השונים	דוברים למשפט	סה"כ מבעים שהוקלטו	משפט לדובר
דיאלקטלי	2	630	1,260	2
דחוס / עשיר	450	7	3,150	5
מגוון	1,890	1	1,890	3
סה"כ	2,342		6,300	10

[3] המילה "שיבולת" משמשת בקרב הבלשנים לתאר אופן הגייה האופייני לדיאלקט מסוים. המילה לקוחה מהפסוק בשופטים יב שבו מסופר כיצד אופן ההגייה של המילה הסגיר את מוצאם של דובריו "וַיֹּאמְרוּ לוֹ אֱמָר-נָא שִׁבֹּלֶת וַיֹּאמֶר סִבֹּלֶת ..." (שופטים יב', ו).

[4] משפטים דחוסים פונטית נקראים בספרות המחקרית גם משפטים *עשירים פונטית* (phonetically rich).

שיטת המחקר

לצורך הכנת בסיס נתונים לאימון של מנוע זיהוי בעברית, החלטנו לא לנסח משפטים מובנים מבחינה פונטית (בדומה למה שנעשה בשני המשפטים הדיאלקטליים בפרויקט TIMIT), אלא לבחור משפטים בשיטה מונחית-נתונים, על ידי תהליך בחירה לולאתי (iterative), המסתמך על שכיחויות צלילי הדיבור ורצפי הצלילים, כך שבסופו של דבר המשפטים שישמשו להקלטות יענו על שתי המטרות: הם גם יהיו דחוסים (עשירים) פונטית, וגם יהיו מגוונים ומבוססים על שימוש סטנדרטי בקורפוס כתוב. היעד של כמות המשפטים עליו הוחלט הוא של 50,000 משפטים שיוקראו על ידי 1,000 דוברים (כל דובר יקרא 50 משפטים).

קורפוס

במחקר הנוכחי בוצע התהליך של יצירת משפטים דחוסים פונטית, לא בצורה אקטיבית, כלומר לא באמצעות ניסוח יצירתי שלהם, אלא בתהליך מונחה-נתונים, שאִפשר את מיצויים מתוך קורפוס ויקיפדיה של "מיל"ה: מרכז ידע לתקשוב בשפה העברית" (מיל"ה; MILA ;Itai & Wintner 2008), המבוסס על מאמרים מהאנציקלופדיה המקוונת של ויקיפדיה בעברית שהורדו ותויגו בשנת 2010. הקורפוס מכיל מעל 130 מיליון תמניות (tokens) (קרוב ל-2 מיליון מילים תבניות(types)). הטקסטים שהורדו היו בפורמט XML (סכמות ה-XML מצייתות לסטנדרטים של מיל"ה) לאחר טוקניזציה (איור 1). המונח "משפט" במחקר הנוכחי הוא על פי סכמת הקידוד של אתר מיל"ה של טקסט המתחיל ב-<sentence> ומסתיים ב-</sentence>. באיור 1 ניתן לראות שהמשפט המודגם מסתיים בסימן הפיסוק – נקודה: "."=surface transliterated="."/></sentence>.

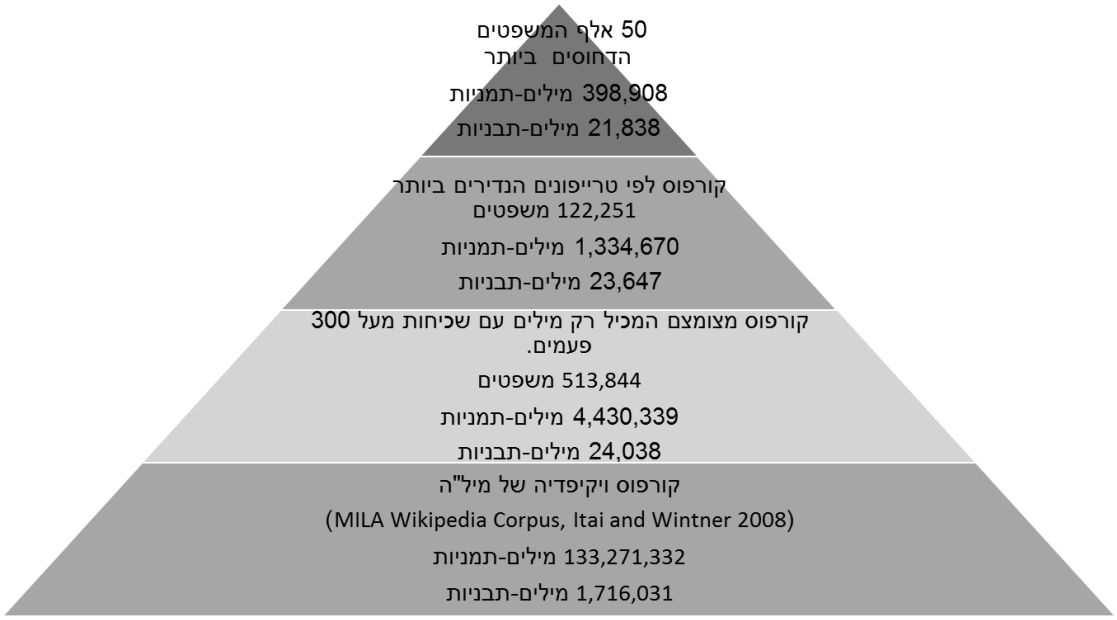
איור מס' 1
דוגמה למשפט "כל נגיף בנוי מצבר מולקולות ביולוגיות, חומר תורשתי (דנ"א או רנ"א, חד-גדילי או דו-גדילי)" בפורמט XML

```
<sentence id="3"><token id="1" surface "כל"=transliterated="kl"/><token id="2" surface "נגיף"=
transliterated="ngip"/><token id="3" surface "בנוי"=transliterated="bnwi"/><token id="4" surface "מצבר"=
transliterated="mcbr"/><token id="5" surface "מולקולות"=transliterated="mwlqwlwt"/><token id="6"
surface "ביולוגיות"=transliterated="biwlwgiwt"/><token id="7" surface="," transliterated=","/><token id="8"
surface "חומר"=transliterated="xwmr"/><token id="9" surface "תורשתי"=transliterated="twreti"/><token id="10"
surface="(" transliterated="("/><token id="11" surface&נד"=quot "א;transliterated="dn&amp;quot;a"/><token
id="12" surface "או"=transliterated="aw"/><token id="13" surface&נר"=quot "א;
transliterated="rn&amp;quot;a"/><token id="14" surface="," transliterated=","/><token id="15" surface "חד"=
transliterated="xd"/><token id="16" surface="-" transliterated="-"/><token id="17" surface "גדילי"=
transliterated="gdili"/><token id="18" surface "או"=transliterated="aw"/><token id="19" surface "דו"=
transliterated="dw"/><token id="20" surface="-" transliterated="-"/><token id="21" surface "גדילי"=
transliterated="gdili"/><token id="22" surface=")" transliterated=")"/><token id="23" surface="."
transliterated="."/></sentence>
```

בשלב הראשוני היה צורך לחלץ את רצף המילים באורתוגרפיה עברית מהטקסטים שהורדו בפורמט XML. לאחר מכן, לאור מטרות הפרויקט לייצר כמות מסוימת של משפטים להקלטות, הוחלט להמשיך ולנתח רק את 500 אלף המשפטים שמכילים את המילים המופיעות יותר מ-300 פעמים בקורפוס

(סף שרירותי שנועד לצמצם את כמות המשפטים). כך נוצר תת-קורפוס ויקיפדיה שכלל 513,844 משפטים, 4,430,339 מילים תמניות (כ-24,038 מילים תבניות) (תרשים 1).

תרשים מס׳ 1
תהליך מיצוי המשפטים להקלטות



שיטת התעתיק ויצירת הלקסיקון

כל המשפטים מתת-הקורפוס עברו תעתוק (transcription) על פי לקסיקון מוכן מראש, כלומר נערכה המרה של הכתיב האורתוגרפי של המשפט לרצף הפונמות שמייצגות אותו, על פי לקסיקון מוכן מראש ועל פי תעתיק אחד. אמנם, כפי שניתן להתרשם מאיור 1 לעיל, פורמט ה-XML מכיל טרנסליטרציה (transliteration) של המילים בעברית. אולם טרנסליטרציה היא המרה חד-חד ערכית של הגרפמה (האות הכתובה) לגרפמה לועזית שנבחרה לייצג אותה, ומכאן שאינה משמשת ייצוג של אופן ההגייה של המילה. התעתיק כלל שימוש ב-31 סימנים המייצגים את 31 הפונמות בעברית המדוברת בישראל (Laufer, 1990; אדם, 2014), והן: 22 העיצורים, כולל ייצוג העיצורים הגרוניים והעיצור הסדקי (ʔ, ʕ, ħ); שלושת המחוככים ([ʧ], [ts], [ʤ]); ו-5 התנועות, כפי שמתועד ב-IPA (Laufer, 1990), וכן העיצור המקורב [w] (ראו לוח 3).

הייצוג הפונמי הומר אוטומטית מהייצוג האורתוגרפי על פי טבלת המרה ("לקסיקון" בעגה של חוקרי טכנולוגיות דיבור) שהוכנה מראש ושכללה שתי עמודות: עמודת המילה באורתוגרפיה עברית ועמודת התעתיק על פי סימני 31 הפונמות (לוח 3). יש לציין כי התעתיק משקף את ההגייה של המילה כפי שהמתעתק שיער כי היתה נאמרת בהגייה בקצב הכתבה. למשל, המבע "אזהרת גניבה", תועתק /ʔazharat gneva/ על פי שתי המילים "אזהרת" ו-"גניבה" שהופיעו בשורות נפרדות בלקסיקון, כפי שמודגם ב-(1).

(1) דוגמה לשורות בלקסיקון

אזהרות	ʔazharot
אזהרת	ʔazharat
אזהרתו	ʔazharato
אזוב	ʔezov
גנטית	genetit
גני	ganej
גניבה	gneva

השיטה ליצירת משפטים דחוסים פונטית

תרשים 1 מציג את תהליך מיצוי הנתונים, החל מהורדת קורפוס "ויקיפדיה" באתר מיל"ה (MILA), וכלה בקביעת 50 אלף המשפטים הדחוסים פונטית, שישמשו להקלטות העתידיות.

על פי Itai and Wintner (2008), קורפוס ויקיפדיה באתר מיל"ה מכיל כ-133,271,332 מילים תמניות ו-1,716,031 מילים תבניות. מהקורפוס המקורי הזה הוצאו, כאמור לעיל, כל המשפטים שמכילים מילים עם פחות מ-300 היקרויות, ונוצר בסיס נתונים מצומצם עם 513,844 משפטים, מעל 4 מיליון מילים תמניות ועם מעל ל-24 אלף מילים תבניות (תרשים 1). בשלב זה הומר הכתיב האורתוגרפי לייצוג פונמי, כמתואר לעיל.

ניתוח שכיחות פונמות ורצפי פונמות[5]

לאחר מכן החל התהליך האיטרטיבי, של בחירת משפטים שמכילים את המספר הרב ביותר של פונמות-תבנית (types), כלומר לא לפי כמות הפונמות, פונמות-תמנית (tokens), אלא לפי גיוון פונמיאלי. התהליך הסתיים לאחר שהתקבל ייצוג של כל הפונמות בשפה. כל התהליך הזה חזר על עצמו כדי לקבל סף מינימום של 700 הופעות מכל פונמה. בשלב זה עדיין היתה כמות גבוהה מדי של משפטים, ולפיכך, בשלב הבא נותחו הנתונים של סביבת הפונמות ברמה של 2 פונמות צמודות (דיפון) עד 3 פונמות צמודות (טריפון). הרצפים שהתקבלו כוללים את שהתקבל ברמת המשפט, כלומר בין-מילים ובתוך-מילים. דוגמה למשפט מתוך בסיס הנתונים עם הטריפון /a d ʔ/ מופיעה ב-(2). הטריפון נמצא ברצף "עד" שבמילה "העדרות":

(2) "טענות מטענות שונות, למשל של העדרות מהארץ."

כדי לייצג את הטריפונים הנדירים ביותר (ברמת המשפט), הוחלט לקחת רק את המשפטים שמכילים טריפונים עם היקרות של בין 99-1. תהליך דומה של מיצוי משפטים בעלי ייצוג טריפונים עשיר מתואר אצל Mendonça et al. (2014).[6] ההנחה היתה כי טריפונים פחות נדירים ממילא צפוי שיופיעו במשפטים האלה. כך נותרו 122,251 משפטים (תרשים 1). כדי לצמצם עוד את כמות המשפטים, בוצע

[5] מכאן ואילך ייעשה השימוש במונח "פונמה" עבור הייצוג שלהן בתעתוקי המשפטים.
[6] שימו לב שפרסום זה הופיע כשנתיים לאחר סיום המחקר הנוכחי.

חישוב יחס בין כמות הטריפונים השונים וכמות הפונמות, ונבחרו המשפטים עם היחס הגבוה ביותר (בהנחה שככל שכמות הטריפונים גבוהה יותר וכמות הפונמות במשפט גבוהה יותר – המשפט עשיר יותר). בסופו של התהליך מוצו 50 אלף משפטים ממופים ודחוסים מבחינה פונטית, שהם גם מגוונים ומבוססים על קורפוס של שימוש טבעי בשפה הכתובה (קצה הפירמידה בתרשים 1).

ממצאים

להלן יוצגו ממצאים כמותיים לגבי מצאי הפונמות ושכיחותן בקורפוס של 50 אלף משפטים מתועתקים שנבחרו לבסיס הנתונים להקלטות, ולגבי רצפי הפונמות ברמת המשפט.

שכיחות הפונמות

לוח 3 מציג את שכיחות כל אחת מ-31 הפונמות שיוצגו בתעתיקים של 50 אלף המשפטים שנבחרו להקלטות. בלוח ניתן לראות כי התנועה /a/ היא הפונמה השכיחה ביותר בשפה ואילו הפונמה השאולה /ʒ/ (ובייצוג אורתוגרפי: ז'), יחד עם יתר 3 הפונמות השאולות, הן הפונמות הנדירות ביותר. את הנתונים בלוח הזה ניתן להשוות לטבלת שכיחות הפונמות בקורפוס של דיבור ספונטני בעברית (ראו נספח). שתי העמודות השמאליות ביותר בלוח 3 מציגות נתונים על תפוצת הפונמות האלה ב-451 שפות העולם, כפי שתועדו על ידי Maddieson (1990) ועל פי הנתונים המופיעים ב-UPSID (The UCLA Phonological Segment Inventory Database)[7].

לוחות 4 ו-5 מציגים את עשרת הדיפונים והטריפונים השכיחים ביותר שנמצאו בקורפוס, ברמת המשפט. קרוב לוודאי שרובם הם הרצפים השכיחים ביותר גם ברמת המילה, ואולם בולט בהקשר זה הטריפון השכיח ביותר /t h a/, שמייצג גם את הרצף המצוי בין מילים "את ה...". לעומת עשרת הדיפונים הנפוצים ביותר, אחד מהדיפונים היחידאיים (singletones) הוא הרצף /ʒ p/, במשפט "הטקסט במקורו מהערך ז'ורז' פרק" /hatekst bemekoro mehaerex ʒorʒ perek/. לעומת עשרת הטריפונים הנפוצים ביותר, אחד מהטריפונים היחידאיים (singletones) הוא הרצף /z v k/, במשפט "מצודת ציון (שיר השירים ז ו) כמין צבע אדומה" /metsudat tsiyun ʃir haʃirim z v kemin tseva ʔaduma/.

מעניינת ההשוואה של נתוני השכיחות של הדיפונים והטריפונים שתוארו לעיל ושלקוחים מבסיס נתונים טקסטואלי, לנתוני השכיחות בבסיס נתונים של קורפוס דיבור ספונטני (מעמ"ד; Silber-Varod, 2013). הדיפונים השכיחים ביותר שנמצאו בדיבור ספונטני הם: /h a/ ו-/m a/ שמופיעים, כל אחד, מעל 2,000 פעמים. אחד הדיפונים היחידאיים הוא /z s/ שנמצא במבע "אז שים את זה פה אני מיד" /az sim et ze po ani miyad/. הטריפון השכיח ביותר בקורפוס הספונטני הוא /a n i/ (שבלוח 5 מופיע במקום השביעי), ואילו אחד היחידאיים /z z o/, שנמצא במבע "אז זאת אומרת כל האנשים מלמעלה יודעים" /az zot omeret kol haanaʃim milemala yodim/.

[7] המידע נלקח מהאתר: http://web.phonetik.uni-frankfurt.de/upsid_info.html, שבו יש ממשק נוח למשתמש של הנתונים של UPSID.

לוח מס׳ 3

שכיחות 31 הפונמות בעברית כתובה בבסיס הנתונים של 50 אלף משפטים מתועתקים (כ-2,205,683 היקרויות של פונמות)

#	הפונמה	היקרויות	שכיחות יחסית (%)	קיימת ב-# שפות (על פי UPSID)	% מתוך 451 שפות (על פי UPSID)
1	התנועה a	361,854	16	392	86.9
2	התנועה e	223,805	10	169	37.5[8]
3	התנועה i	163,572	7	393	87.1
4	m	131,839	6	425	94.2
5	t	113,334	5	181	40.1
6	l	113,104	5	174	38.6
7	התנועה o	108,455	5	131	29.0
8	r	97,732	4	95	21.1
9	h	80,727	4	279	61.9
10	n	78,013	4	202	44.8
11	ʃ (ש)	68,214	3	187	41.5
12	k	63,891	3	403	89.4
13	התנועה u	61,397	3	369	81.8
14	y	60,957	3	378	83.81
15	ʔ (א)	59,973	3	216	47.9
16	b	52,258	2	287	63.6
17	v	48,786	2	95	21.06
18	ħ (ח)	47,988	2	19	4.21
19	d	44,269	2	120	26.61
20	s	37,115	2	196	43.5
21	ʕ (ע)	34,622	2	10	2.22
22	x (כ)	33,536	2	44	9.76
23	ts (צ)	28,119	1	62	13.75
24	f	23,855	1	180	39.9
25	z	20,425	1	62	13.75
26	g	19,421	1	253	56.1
27	p	18,512	1	375	83.2
28	dʒ (ג׳)	4,461	0.20	113	25.06
29	w (וו)	2,601	0.12	332	73.6
30	tʃ (צ׳)	2,494	0.11	188	41.7
31	ʒ (ז׳)	354	0.02	61	13.53

[8] כרפרנס נבחרה התנועה הקדמית-אמצעית בלתי-מעוגלת (mid front unrounded vowel) /e/. על הגיוון בהגיית תנועה זאת בעברית, ראו: Silber-Varod & Amir (2016).

לוח מס׳ 4
עשרת הדיפונים הנפוצים ביותר (ברמת המשפט), בבסיס הנתונים של 50 אלף משפטים מתועתקים

#	דיפון	היקרויות
1	h a	58,096
2	i m	39,810
3	a t	32,519
4	a r	31,111
5	a m	30,685
6	ʔ a	30,536
7	a l	26,990
8	b e	25,358
9	l e	24,710
10	a n	24,291

לוח מס׳ 5
עשרת הטריפונים הנפוצים ביותר (ברמת המשפט)

#	טריפון	היקרויות
1	t h a	12,236
2	ʕ a t	8,955
3	ʃ e l	8,331
4	e r e	7,856
5	a h a	7,809
6	h a m	7,335
7	a n i	7,038
8	a ʔ a	6,772
9	a ħ a	6,658
10	a m a	6,646

דיון

מחקר יישומי זה נועד בראש ובראשונה להפיק רשימת 50,000 משפטים עשירים-פונטית ומגוונים שישמשו להקלטות דוברי עברית. מכיוון שההקלטות נועדו לשמש תשתית לאימון צלילי הדיבור עבור מנוע זיהוי דיבור, נדרש היה להפיק משפטים שבהם יהיה ייצוג של כל אחת מ-31 הפונמות בשפה בכמות מספקת, ובכל ההקשרים הפונטיים המתממשים בשפה.

הדרישה הראשונה בנוגע לייצוג הפונמות תוארה במאמר זה והממצאים מציגים את השכיחות של כל אחת מ-31 הפונמות בשפה ואת סדר השכיחות מהפונמה השכיחה ביותר /a/ לפונמה הנדירה ביותר /ʒ/. עוד נמצא כי הממצאים אינם תואמים באופן מלא את היררכיית המסומננות, במיוחד לאור העבודה שעיצור [p] נמצא במיקום ה-27 ברמת השכיחות ואילו עיצורים מסומננים יותר, כמו [x] ו-[k] נמצאו במיקומים גבוהים יותר (22 ו-12, בהתאמה). הנתון על שכיחותו הנמוכה של העיצור [p] מפתיע גם לאור שכיחותו הגבוהה בשפות העולם (כ-83%; ראו לוח 3).

כמו כן, ניתן לראות כי הפונמות השאולות הן הנדירות ביותר בשפה הכתובה כפי שבאה לידי ביטוי בקורפוס ויקיפדיה 2010 (מיל״ה). בנוסף, בנספח למאמר זה מופיעה טבלת שכיחויות הפונמות בשפה המדוברת הספונטנית (מעמ״ד; Silber-Varod, 2013). אמנם סדר שכיחות הפונמות בנספח אינו תואם

בדיוק לסדר שמופיע בלוח 3, אולם הוא אינו שונה לחלוטין ואפשר לייחס את זה גם להיקף השונה מאוד של שני הקורפורה. מכל מקום, נדרש מחקר מקיף יותר של שכיחות הפונמות בשפה המדוברת הספונטנית שגם יעמוד על ההבדלים בתפוצת הפונמות ורצפי הפונמות בין השפה הכתובה והשפה המדוברת.

הדרישה השנייה בנוגע להקשרים הפונטיים היא עמומה במובן זה ש"הקשרים פונטיים" יכולים, מחד, להיות הקשרים ברמת ההברה, ולכן כאלה שמצייתים לחוקי השפה, ובהם היררכיית הצליליות וחוקי הפונוטקטיקה. לעומת זאת, "הקשרים פונטיים", או "סביבות פונטיות", ברמת המשפט הם מגוונים, ולכן היה חשוב לאפשר אותם, מתוך מטרה להקליט את המימוש שלהם. במחקר הנוכחי המצב הרצוי היה של ייצוג כל הרצפים האפשריים בשפה בתהליך מונחה-נתונים. בתהליך זה נמצאו גם רצפים נדירים ביותר (ראו סעיף "מגבלות המחקר הנוכחי" בהמשך).

כבר הראנו לעיל בלוחות 4 ו-5, שחלק מהדיפונים והטריפונים שנמצאו במצאי המשפטים הם כאלה שנוצרו בין הברות או בין מילים, ואולם, בדיקה של מצאי הטריפונים העלתה כי לא כל הרצפים שאפשריים בשפה מופיעים בבסיס הנתונים של 50 אלף המשפטים. למשל, הטריפון [p d i] אינו מופיע במשפטים שנאספו, על אף שכן נמצאים הטריפונים pda, pde, pdo, pdu. לוח 6 ממחיש כיצד צרורות עיצורים שכן אפשריים בשפה העברית בעמדת האונסט, אינם מופיעים במשפטים שמוצו בתהליך מונחה-נתונים להקלטות (הסימן + מסמן שהטריפון נמצא, והסימן – שהוא אינו נמצא במשפטים).

לוח מס' 6

טריפונים אפשריים בעברית לעומת מצאי בקורפוס 50 אלף המשפטים. האות V מסמנת את עמדת גרעין ההברה.

טריפונים אפשריים בעברית	גרעין ההברה a	גרעין ההברה e	גרעין ההברה i	גרעין ההברה o	גרעין ההברה u
p d V	+	+	-	-	+
p g V	+	-	+	-	-
p n V	+	+	+	+	-
p z V	+	+	-	-	-
p ʃ V	+	+	+	-	+
p x V	+	-	-	-	-
p ts V	+	-	+	-	+
b x V	-	+	+	-	-
t x V	+	+	+	+	+
d ħ V	+	+	+	+	+
k z V	+	+	+	+	-
k x V	+	-	+	+	+
g l V	+	+	+	+	+

כך, התהליך של מיצוי המשפטים מקורפוס של 100 מיליון מילים חשף דיפונים וטריפונים מגוונים, כלומר הקשרים מגוונים לכל הפונמות, מעבר לרמת ההברה והמילה, כגון /w w o/ ברצף המילים "וואוו וואוו", או הטריפון /r f g/ שהתקבל פעם אחת בקורפוס, בתעתיק של המשפט: "שֶׁרְף גם שימש כמאמן הנבחרת הלאומית של ישראל בכדורסל גברים.". כמו כן, ניתוח הנתונים העלה שנמצאו רצפים ברמת המשפט שגם אפשריים ברמת ההברה (למשל, /k x e/), אך טרם נוצרה מילה בעברית המכילה אותם. לעומת זאת, רצפים שכן מוכרים ברמת ההברה, כגון /pzu, pzi/ או /pgo/, או שנשמטו במהלך העיבוד משום שהם נדירים יותר (כגון /b x a/), או שפשוט לא הופיעו בקורפוס הנחקר.

מגבלות המחקר הנוכחי

מהתיאור לעיל ניתן לראות שהתהליך של ייצוג פונמות ברמת הפונמה הבודדת, ברמת הדיפון, וברמת הטריפון בשיטה מונחית-נתונים אינו חף מחסרונות. הראשון שבהם הוא בשלב ההמרה האוטומטית מכתיב אורתוגרפי לתעתיק באמצעות הלקסיקון. תהליך זה לא אפשר בחירת התעתיק המתאים למילה הנמצאת בתוך משפט, אלא הבחירה היתה של תעתיק אחד למילים שונות עם אורתוגרפיה זהה (הומוגרפים). למשל, רצף הגרפמות "ספר" תועתק בכל המקרים כ-/s e f e r/ (רצף זה של 5 פונמות מופיע 551 פעמים). אין זה אומר שאין ייצוג לרצף /s a f a r/ (רצף זה מופיע 8 פעמים) או "סְפָר" /s f a r/ (רצף זה מופיע 226 פעמים), אך במקרים מסוימים הדבר אומר שאופן התעתוק אינו זהה לאופן ההקראה הצפוי.

חיסרון נוסף הוא שהתעתיק ברמת המשפט התעלם מתהליכים פונוטקטיים של נשילת עיצורים ותנועות המתרחשים בדיבור טבעי. למשל, ברצף "אז זאת אומרת...", נכלל אוטומטית הדיפון /zz/, דבר שלא מתרחש בדיבור טבעי, שבו יש התמזגות לעיצור אחד. דוגמה נוספת היא של ייצוג קבוע למילה "המנהל" כ-/hamenahel/, על אף שצפויה הגייה של /hamnahel/ (בולוצקי, 2002). על חיסרון זה אפשר להתגבר אם בעת ההקלטות מבקשים מהדוברים להגות את המשפט בקצב הכתבה ולהקפיד על הגיית כל מילה בנפרד או לחילופין, מקפידים לתעתק את המבע תעתוק פונטי צר, כפי שהוא נהגה בפועל, ולא על פי מילון מוכן מראש. לבסוף, בתהליך מיצוי המשפטים "אבדו" רצפים אפשריים בשפה.

סיכום

המחקר הנוכחי הוא יישומי במהותו, אולם הוא תורם גם לחקר הפונולוגיה בעברית המדוברת, בכמה היבטים. הראשון בחקירת תפוצת הפונמות ותפוצת הרצפים (דיפונים וטריפונים) ברמת המשפט בשפה הכתובה; השני, בהעלאת המודעות לחשיבות הסביבות הפונטיות ברמות הפרוזודיות שמעבר לרמת ההברה – ברמת המילה והמבע. בכך תורם המחקר הנוכחי לחשיפת הצירופים האפשריים בעברית ותפוצתם: "כדי להשלים את המחקר [על חוקי צירופי הפונמות, חוקים פונוטקטיים, הערת המחברים] יש להכין מאגר גדול של עברית בתעתיק פונטי ופונמי ולגלות על פיו לא רק את הצירופים האפשריים, אלא גם את תפוצתם. מידע כזה הוא סעיף בדקדוק לשוננו ..." (לאופר תשנ"א, עמ' 190).מחקר עתידי ראוי שיבחן את שילוב כל הרצפים האפשריים בשפה ברמת ההברה ואת תפוצתם. כמו כן, מן הנתונים שהוצגו עולה החשיבות של תעתיקים המשקפים את אופן ההגייה בפועל. אלה, כידוע, דורשים משאבים ומיומנות גבוהה של פונטיקאים. לבסוף, ההקלטות שיתבצעו בעתיד על בסיס המחקר הזה, יכולות לשמש להקמת ספריית

צלילי הדיבור בעברית, שבה יוקלטו כל ההגיות של כל פונמה וסביבותיה במאגר של הקלטות, שיתעד את מאפייני ההגייה בפועל של דוברי העברית המדוברת וכדי ללמוד עוד על השפה הסובבת אותנו.

נספח

מצאי 28 הפונמות בעברית (כ-109,880 היקרויות של פונמות) בקורפוס דיבור ספונטני (מעמ"ד ; Silber-Varod, 2013)[9]

#	הפונמה	היקרויות	שכיחות יחסית (%)
1	a	19,298	18
2	e	13,690	12[10]
3	i	8,446	8
4	m	6,710	6
5	l	6,690	6
6	o	6,444	6
7	t	5,894	5
8	n	4,320	4
9	r	4,028	4
10	h	3,880	4
11	ʃ	3,726	3
12	k	3,426	3
13	x	3,339	3
14	y	3,055	3
15	v	2,579	2
16	u	2,396	2
17	z	2,311	2
18	b	2,310	2
19	d	1,934	2
20	s	1,594	1
21	ts	1,151	1
22	g	906	0.82
23	f	899	0.82
24	p	695	0.63
25	w	76	0.07
26	tʃ	44	0.04
27	dʒ	32	0.03
28	ʒ	7	0.01

[9] בסיס הנתונים כלל 31 אלף מילים.

[10] הפונמה /e/ כוללת גם תעתיק של היסוסים בדיבור.

מקורות

אדם, ג. (2014). *מבוא לבלשנות תאורטית, חלק ד*. רעננה: הוצאת האוניברסיטה הפתוחה.

בולוצקי, ש. (2002). שונות פונולוגית ומורפולוגיה בעברית המדוברת. בתוך ש. יזרעאל (עורך), *תעודה, יח - מדברים עברית: לחקר הלשון המדוברת והשונות הלשונית בישראל*. (עמ' 239-278). תל-אביב: אוניברסיטת תל-אביב.

זילבר-ורוד, ו., לטין, מ., ומויאל, ע. (2013). מאגר הפונמות בעברית – פרויקט שתחילתו במאה מיליון מילים ותכליתו 31 פונמות. הוצג במפגש ה-29 של החוג הישראלי לבלשנות ע"ש חיים רוזן. המכללה האקדמית צפת. 4 בפברואר 2013.

לאופר, א. (תשנ"א) צירופי פונמות – פונוטקטיקה. בתוך: מ. גושן-גוטשטיין, ש. מורג, ש. קוגוט (עורכים). *שי לחיים רבין - אסופת מחקרי לשון לכבודו במלאת לו שבעים וחמש*. סדרת ספרי מחקר במדעי היהדות (עמ' 179-193). ירושלים: אקדמון.

מיל"ה: מרכז ידע לתקשוב בשפה העברית. אתר אינטרנט: ttp://www.mila.cs.technion.ac.il/heb/index.html

מאגר העברית המדוברת בישראל (מעמ"ד). אתר אינטרנט: < http://cosih.com/index.html >

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

Garofolo, J. S., Lamel, L. F., Fisher, W. M., Fiscus, J. G., Pallett, D. S., Dahlgren, N. L., & Zue, V.R. (1993). *TIMIT Acoustic-Phonetic Continuous Speech Corpus*. Available at: http://www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC93S1

Gibbon, D., Moore, R., & Winski, R. (Eds.). (1997). *Handbook of standards and resources for spoken language systems*. Berlin: Walter de Gruyter.

Hay, J., & Bauer, L. (2007). Phoneme inventory size and population size. *Language*, *83*(2), 388-400.

Lamel, L. F., Kassel, R. H., & Seneff, S. (1989). Speech database development: Design and analysis of the acoustic-phonetic corpus. *ESCA Tutorial and Research Workshop on Speech Input/Output Assessment and Speech Databases, Vol. 2* (pp. 161-170), Noordwijkerhout, The Netherlands.

Laufer, A. (1990). Hebrew. *Journal of the International Phonetic Association*, *20*(2), 40-43. doi:10.1017/S0025100300004278

Maddieson, I. (1984). *Patterns of Sounds*. Cambridge: Cambridge University Press.

Maddieson, I., & Precoda, K. (1990). Updating UPSID. *UCLA Working Papers in Phonetics 74*, 104-111.

Mendonça, G., Candeias, S., Perdigão, F., Shulby, C., Toniazzo, R., Klautau, A., & Aluísio, S. (2014). A method for the extraction of phonetically-rich triphone sentences. *2014 International Telecommunications Symposium (ITS)* (pp. 1-5). IEEE.

MILA: Knowledge Center for Processing Hebrew. (website). Available at: http://www.mila.cs.technion.ac.il/heb/index.html

Silber-Varod, V. & Amir, N. (2016). Formant analysis of the mid-front vowel as realized in hesitation disfluencies in Hebrew. *Phonetician*. *113*, 49-60.

Silber-Varod, V. (2013). *The SpeeCHain Perspective: Form and Function of Prosodic Boundary Tones in Spontaneous Spoken Hebrew*. Saarbrücken, Germany: LAP Lambert Academic Publishing.

Schmid, S. (2012). Phonological typology, rhythm types and the phonetics-phonology interface. A methodological overview and three case studies on Italo-Romance dialects. In: A. Ender, A. Leemann, & B. Wälchli (Eds.), *Methods in contemporary linguistics* (pp. 45-68). Berlin: De Gruyter Mouton. ISBN 978-3-11-028466-9.

TIMIT – Acoustic-Phonetic Continuous Speech Corpus. Available at: https://catalog.ldc.upenn.edu/LDC93S1

UPSID – The UCLA Phonological Segment Inventory Database. Available at: http://www.linguistics.ucla.edu/faciliti/sales/software.htm#upsid

Wolf, L., & Tobin, Y. (2011). Tendential Strategies in Consonant Inventories across Languages According to the Theory of Phonology as Human Behavior. In W. S. Lee & E. Zee (Eds.), *Proceedings of the 17th International Congress of Phonetic Sciences* (pp. 2141-2144).

Zue, V., Seneff, S., & Glass, J. (1990). Speech database development at MIT: TIMIT and beyond. *Speech Communication*, *9*(4), 351-356.