

## מאגר העברית המדוברת בישראל (מעמ"ד) שלב א': בדיקה טרומית – דו"ח ראשוני

שלמה יזרעאלי<sup>1</sup>

לפרופסור יעקב כרסולילה, עמית וידי, חבר צוות מעמ"ד,  
שוכות גדולה לו בחקר העברית המדוברת בישראל

מאגר העברית המדוברת בישראל (מעמ"ד) הוא פרויקט שכא למלא חסר רב שנים במחקר הלשון והחברה בישראל. מטרתו:  
א. כינון מאגר של העברית הישראלית המדוברת כתשתית למחקר שיטתי. מחקר המאגר יקוי מיגוון רחב של נושאים הקשורים בשפה העברית ובמתודולוגיה הכללית של חקר הלשון המתבסס על מאגרי לשון.  
ב. הפצת המאגר לציבור במולטימדיה ובדפוס. ההפצה כאמצעים אלקטרוניים – CD-ROM, DVD-ROM והאינטרנט – תיעשה כך שהקלטות ותמלילים מוצגים במקביל ובשילוב דרכי תיעוד וניתוח נוספות.

המאגר יהיה נגיש ופתוח לכול ולכל צורך.  
מאגר העברית המדוברת בישראל (מעמ"ד) יכלול את המאפיינים האלה:

1. הקלטות קול דיגיטליות
  2. מבחר הקלטות דיגיטליות בווידיאו
  3. תמלילי כל הטקסטים שבמאגר
  4. תעתיק פונטי של קטעים נבחרים
  5. ניתוח מורפולוגי רציף של קטעים נבחרים
  6. תרגום לאנגלית של קטעים נבחרים
- מעמ"ד יכיל כחמישה מיליון מילה. הוא יורכב מאלף תאים בני חמשת אלפים מילה כל אחד, ומידגם מייצג ממנו, המישה אחוזים מהתאים, וקלט בווידיאו. "תא" הוא היחידה הסוציולווגיסטית הבסיסית במאגר. התא הוא קטע דיבור מוקלט בן חמשת אלפים

1 פרויקט מעמ"ד הוא פרויקט מורכב, שאי אפשר לו להתבצע על ידי חוקר יחיד. צוות כינון המאגר: שלמה יזרעאלי (ראש התוכנית וחוקר ראשי); בנימין הרי (חוקר ראשי); ג'ון רו באוא (אנליסט המאגר); מירה אריאל (חקר השיח ופרגמטיקה); גיורא רחב (סוציולוגיה וסטטיסטיקה). צוות ייעצים: אליעזר בן רפאל (סוציולווגיסטיקה – היבטים סוציולווגיים); יעקב כרסולילה (סוציולווגיסטיקה – היבטים בלשניים); אוטו יטורו (תיעתוק, פונולוגיה ודיאלקטולוגיה); שמואל בולוצקי (פונולוגיה, מורפולוגיה); ג'פרי כאן (תחביר); אילנה שוהמי (חינוך לשוני). לחכנון המחקר המקדים ולהוצאתו לפועל עסקו עמי בעיקר גיורא רחב ובנימין הרי. בניבוש השאלון הסוציולווגיסטי עזרה גם אילנה שוהמי. משה כסלו שימש בנאמנות כעוזר טכני. תודתי להם לכול חברי הצוות, שכלי פעילותם ומסירותם דבר לא יוכל לצאת לפועל.

מילה של טקסט רציף ומלוכד או כמה קטעים שכוללים יחדיו חמשת אלפים מילה. המאגר ישקף מצב סינכרוני של העברית המדוברת בישראל. היות שמדובר בשפה טבעית, ובמיוחד בעברית הישראלית המשתנה בתכיפות, יש להשלים את מעמ"ד בפך זמן קצר ככל האפשר.

מעמ"ד יורכב משני תת-מאגרים: מאגר ראשי ומאגר משלים. המאגר הראשי יהווה את חלק הארי של המאגר כולו, ויכלול כתשעים אחוז מכלל הנתונים. המאגר המשלים יכלול כעשרה אחוזים מהנתונים. המאגר הראשי יורכב מתאים שייבנו על פי קריטריונים דמוגראפיים משולבים בקריטריונים לבחינת הקשרי השיח. הנתונים במאגר המשלים ייערכו בשני תת-מאגרים שונים: תת-מאגר אחד ייעד להשלמת המאגר הראשי וישותת בעיקרו על קריטריונים דמוגראפיים. תת-מאגר זה ידגום בדגימה לא מאוזנת קבוצות אוכלוסין אשר ייצוגן יהיה חסר במאגר הראשי. תת-מאגר השני וישותת בעיקרו על קריטריונים של הקשרי שיח, ויכיל דגימות מקסטים מאמצעי התקשורת, מן הכנסת ומבתי המשפט. כל תת-מאגר מן המאגר המשלים יורכב מחמישים תאים, דהיינו המישה אחוזים מן המאגר כולו.

הגישה שנקטה על ידינו לכינון המאגר הראשי, הוא המאגר המייצג, היא גישה תלוית תרבות. מעמ"ד שואף לגשר בין אינספור חלופות השיח הנהוגות בקרב קהילת הדוברים של העברית הישראלית לבין ייצוגיותה במאגר על ידי אפיון השונות במונחים דמוגראפיים והקשריים גם יחד. זהו ניסיון ראשון וייחודי לכונן מאגר מייצג על שני ציורי משתנים, דמוגראפיים ותלויי הקשר, בהתאם לקריטריונים סטטיסטיים ואנליטיים. בחירת האינפורמנטים להקלטות המאגר הראשי תבצע על ידי דגימה אקראית של אוכלוסיית ישראל, במטרה לשקף את המבנה החברתי של קהילת דוברי העברית בישראל. פילוח המאגר לצרכים אנליטיים ייקבע על פי קריטריונים מוגדרים, אף כי כל הנתונים הסוציולinguיסטיים של האינפורמנטים שיוקלטו יעמדו לרשות המשתמשים במעמ"ד. הנחת העבודה שלנו מתבססת על שלושה קריטריונים דמוגראפיים שנראים לנו החשובים ביותר ביצירת השונות הלשונית בישראל: (1) מקום לידה, ארץ מוצא המשפחה, ועדה או רח; (2) גיל; (3) השכלה. הנחת העבודה שלנו לניתוח הקשרי שיח מתבססת על חמישה משתנים: יחסים בינאישיים, מבנה השיח, ונושא השיח כמשתנים עיקריים, ולצידם מספר המשתתפים בשיח וערוץ התקשורת. עריין לא נערך מחקר מקיף בחברה הישראלית לבדיקת הנתונים הדמוגראפיים והנתונים ברבר הקשרי השיח בשימוש העברית בה. לפיכך, כדי לעצב דגם ראוי למעמ"ד, המחקר וכינון המאגר מתבצעים בשלבים, אשר במהלכם ייבדקו התאמת הנתונים הדמוגראפיים והנתונים תלויי ההקשר כפי שהוצעו לעיל. כל שלב יאפשר לנו לאשש את השלבים הקודמים, ויסייע בכינון לגבש את צורתו הסופית של המאגר בשלמותו.

שלב המחקר, הלימוד והתכנון של מעמ"ד ארך שלוש שנים. בתקופה זו נלמדו אפשרויות כינון מאגרי לשון ומצאי המאגרים בעולם כיום; נבדקו דרכי אגירה ודרכי תצוגה; נחקרו הדרכים התיאורטיות לכינון מאגר לשון; וגובשה תוכנית לכינון מעמ"ד. כפי שנוכר לעיל, התוכנית לכינון מעמ"ד היא תוכנית חדשנית וייחודית. ייחודיות

מעמ"ד אינו רק בהיותו מאגר של עברית מדוברת, אלא בהיותו מאגר מייצג לצורות הלשון השונות המדוברות באוכלוסייה. מאמר מדעי תיאורי על המאגר ותכנונו המידגמי הוזמן על ידי כתב-העת *Corpus Linguistics* (IJCL) (Izre'el, Hary and Rahav 2001) (יזרעאל, הרי ורהב תשס"ב); ראה גם: <http://spinoza.tau.ac.il/humanities/semitic/maamad.html>.

כדפים אלה אציג את תוכנית השלב הראשון, הכולל בדיקה טרומית (pilot) וראשיתו של מחקר מקדים (pretest). תוכנית זו יש לראות כיחס לתכנון מעמ"ד ומבנהו. הקורא מוזמן, אפוא, לעיין במאמרים הרשומים לעיל, שם ימצא את פירוט תוכנית מעמ"ד עם הסבר ההנחות שעמדו בפנינו בתכנונו.

מטרות:

1. חקירה מקדימה שתסקור מיגוון מחלופות השיח בקהילת דוברי העברית בישראל;
2. גישושים ראשוניים לקראת לימוד הנושאים הכרוכים בדגימה אקראית של האוכלוסייה;
3. לימוד דרכי השגת ההקלטות ומידע נוסף (נסיבות, סוציולinguיסטי, וכיו"ב);
4. לימוד דרכי השגת הנתונים מקבוצות שונות בחברה הישראלית;
5. לימוד הבעיות הכרוכות בהקלטה ארוכת טווח ורציפה;
6. בדיקת הכלים הטכניים, שיטות הרישום, התימלול, התעתיק וסוגי הטקסט.

הליכי ביצוע:

- א. גיוס אינפורמנטים בדגימה לא סטטיסטית, מכוונת;
- ב. הכנת האינפורמנט לקראת ההקלטות;
- ג. הקלטה ברצף על ידי האינפורמנט עצמו בפרקי זמן משתנים;
- ד. תישאל האינפורמנטים ברבר התנאים והנסיבות של ההקלטות שנעשו;
- ה. ראיון סוציולinguיסטי עם כל אינפורמנט;
- ו. שאלות טכניות לבדיקת דרכי העבודה;
- ז. החתמה על הסכמה לשימוש ולהפצת ההקלטות לצורכי מחקר;
- ח. ארגון ומיון ראשוני של החומר הגולמי;
- ט. הערכת ההקלטות: איכות, כמות החומר, וכיו"ב;
- י. ברירת נתונים: שעה מתוך פרק הזמן שהוקלט;
- יא. מיון החומר, הגדרתו ורישום במסד נתונים;
- יב. תימלול;

- יג. בחירת קטעים להשוויות נוספות: תעתק פונטי, ניתוח מורפולוגי, תרגום לאנגלית;
- יד. ניתוח סקסטואלי לחיבור חלופות הלשון ומיון טיפולוגי של טקסטים לעיון לקסיקלי ודקדוקי;
- טו. הערכת שלב I כמכלול.

**פירוט ולקחים ראשוניים**

א. גיוס אינפורמנטים לצורך הבדיקה הטרומית תוכנן שייעשה בדגימה לא סטטיסטית. בנינו תבנית תאים בסיסית, שעל פיה יש לבחור את המשחתתים במידגם:

גיל	השכלה	אשכנזים	מזרחים	ערבים	אחרים
צעירים	נמוכה/תיכונית				
מבוגרים	נמוכה/תיכונית				
	גבוהה				
	גבוהה				

האינפורמנטים המוקלטים בבדיקה הטרומית נבחרים על פי קריטריונים דומים לקריטריונים האנליטיים שנבחרו כהנחת העבודה לשימוש במעמ"ד: מוצא, גיל והשכלה. מספר חלופות שיח תלויות הקשר יבואו לידי ביטוי תוך כדי הקלטה, ותשומת הלב לחלופות אלה תהיה על פי המצאי שיתקבל. יש לשים לב, עם זאת, שהבדיקה הטרומית אמנם מיועדת להציג שונות לשונית, אך לא לנחתה. משום שמטרת הבדיקה הטרומית היא לסקור מיגוון רחב ככל האפשר מתוך חלופות השיח בקהילת דוברי העברית בישראל, המידגם הזה נבנה מתוך עדיפות לגיוון דמוגרפי שיהיה רחב ככל הניתן במסגרת מצומצמת זו. משום כך הגדרנו צעירים כגילאי 20-27 שאינם בעלי משפחה ומבוגרים כגילאי 50 ומעלה שהם בעלי משפחה. ייצוג על פי מין יתקבל עם מתן הדעת לייצוג שווה במין בתבנית, אף אם לא בהכרח ביחס התפוצה שווים למשתתפים האחרים. כפי שנראה להלן, גיוס האינפורמנטים ואיסוף ההקלטות לבדיקה הטרומית יתבצע על ידי שלושה גופים שונים. כל גוף יגייס שישה עשר אינפורמנטים, על פי הנתונים בטבלה המוצגת: ארבעה יהודים ממוצא אשכנזי, ארבעה יהודים ממוצא מזרחי, ארבעה ערבים וארבעה אחרים על פי הפירוט כדלקמן: גוף א' יגייס ארבעה חיילים; שניים צעירים, חייל וחילת כשרות חובה, ושניים מבוגרים, המשרתים בצבא קבע או בכוחות הביטחון בני למעלה מ-40 ובעלי משפחה. גוף ב' יגייס ארבעה חרדים או דתיים מאוד החיים בקהילה חרדית או דתית, עם ייצוג לאשכנזים וספרדים, לגברים ונשים. גוף ג' יגייס ארבעה עולים חדשים כשגילם והשכלתם תואמים את המיפרט הכללי, שני גברים ושתי נשים.

דגימת אוכלוסי ישראל וגיוס אינפורמנטים אינה משימה של מה בכך, וראי לא לבלשנים (סקירה על דרכי איסוף החומר למחקר לשון טבעית ר' אצל Milroy 1987). משאבי האנוש שצפויים להתרכז בעבודה על מעמ"ד אינם מיומנים ולא יוערו ללימוד, לאירגון ולבנייה של מידגם אוכלוסי מייצג. גם מנקודת ראות כלכלית אין זה ראוי – ואולי אף לא ישים – לרכז משאבי אירגון ומימוש לצורך זה. לפיכך ראינו לנכון להטיל פעולה זו על גוף מיומן יותר ובעל פריסה ארצית. גוף כזה הוא, למשל, חברה למחקר שווקים או מכון לחקר דעת קהל (המאגר הלאומי הבריטי – The National British Corpus – שכרו את שירותיו של המשרד לחקר שווקים The British

<http://info/ox/ac/uk/bnc/what/raa Market Research Bureau; ראה [spok\\_design.html](http://spok_design.html). כדי לבדוק את אפשרות הביצוע של כינון מעמ"ד בעזרת גוף כזה פנינו לשם הבדיקה הטרומית לשלושה גופים העוסקים בסקרי דעת קהל. מן הניסיון שכבר נצבר עם אחד הגופים הללו, נראה כי אכן נוכל להסייך מעל שכמנו את עול איסוף ההקלטות ולהפנות את עיקר משאבינו ואת עיקר מרצנו לעבודה הכלנית ולמחקר הסוציולוגיטי. בשלב זה, המידגם המכוון משלב בתוכו גם ראשית לימודן של דרכי גיוס האינפורמנטים. המידגם המכוון שתכננו אנו יועד לבחירת אינפורמנטים כשיטת "חבר מביא חבר". שיטת גיוס האינפורמנטים שנוקט הגוף הזה משלבת בתוכה גם חיפוש אקראי של אינפורמנטים: הליכה מרלת לרלת באזור נתון וניסיון לגייס אינפורמנטים מתוך קהל לא מוכר.

מכצע של גיוס אינפורמנטים אינו יז לחלוטין לגופים כגון אלה. לא כן שכנוע אורח מן השורה להשתתף במחקר שדורש הקדשת זמן ומאמץ מצדו. משום כך יש לתדרך ולאמן את עובדי מכוני הסקרים הללו. לקח ראשון שנלמד הוא הקושי שבניוס ראשוני של סוקרים לצורך מחקר מסוכך כזה. על האינפורמנטים – ובראש ובראשונה על הסוקרים – להשתכנע בהשיבותו של המחקר. אינפורמנט אייש לא ישתף פעולה. שכנוע כזה לא יוכל לבוא אלא מפי סוקר שמשוכנע הוא עצמו בחשיבות המחקר.

מניסיונו המועט נוכל לשתף את הקורא בשני פכים קטנים: אינפורמנט צעיר אחד אמר שלא יקש תשלום אם ההקלטות לא יועילו למחקר. אינפורמנט אחר, מבוגר, שהיחס תחילה אם ליטול חלק במחקר, בסופו של דבר הלהבה, וכתום פרק ההקלטה אף שלח פתק ביוזמתו, בו נאמר:

נתבקשתי להשתתף בסקר מטעם האוניברסיטה, ואני עושה זאת ברצון! (... אני מאחל לעורכי המחקר הצלחה מלאה!!

מידת ההיענות של האוכלוסייה היא אכן שאלה חשובה. שאלה זו נבדקה בסקרים ראשוניים. סקר פנים-אל-פנים שערך מכוון ב. ג' ולוסיל כהן (אוניברסיטת תל-אביב) באוגוסט-ספטמבר 2001 העמיד את השאלה הזו:

האם תהיה מוכן להשתתף כעתיד במחקר ייחודי שבו תתבקש תמורת תשלום להקליט את פעילותך במשך יומה אחת?

הסקר נערך על 1170 נשאלים, שהיוו מידגם מייצג של האוכלוסייה. אחוז המשיבים בחיוב היה 40%, לעומת 60% צעיינו כי לא יסכימו. במיגור הערבי נשאלו 150 איש, ואחוז ההיענות שם היה 24%. מכוני סקרים מודעים לעובדות אחרות לגבי היענות לסקרי דעת קהל במיגורים שונים. למשל, בעוד שבמיגור היהודי ההיענות גבוהה יותר אצל נשים מאשר אצל גברים, הרי ההיפך הוא הנכון במיגור הערבי, בו ההיענות אצל נשים נמוכה יותר. במיגור הערבי בעיה זו תיווסף לבעיה נוספת הקשורה בשאלת מידת השימוש בעברית. נשים, במיוחד בגילאים הגבוהים ואלה שמבלות רוב עיתותיהן בסביבה דוברת ערבית, לא הדברנה עברית כלל או כמעט כלל. דגימה אקראית במיגור הערבי עולה, אם כן, לייצור בעיה של חוסר בנתונים.

בעיה נוספת שעלתה מגישושים ראשונים לגיוס אינפורמנטים במיגור הערבי הולתה חוסר רצון בולט להשתתף בפרייקט גם מצד גברים. החשש מפני חדייה לפרטיות, חשש שיכול להיות אקוטי במיגורם רחבים באוכלוסייה, ייתכן שיהיה עמוק יותר במיגור הערבי. דבר זה צריך בדיקה. אולם במיגור הערבי עלה בנוסף גם החשש מפני נשיאת מכשיר הקלטה בפומבי, דבר העלול להחשיד את נושאו בשיתוף פעולה עם השלטונות.

שאלת תשומת הלב הניתנת למכשיר ההקלטה על ידי הסוכבים היא שאלה כללית יותר. מכשיר ההקלטה חבוי בתיק מיוחד, ככל המיקרופונים מתחת לבגדים, והמיקרופונים עצמם מתחת לצווארון. כך התכונן, כך האידיאל. מעשית אין הדבר תמיד כך: המיקרופונים עלולים לבצבץ, והתיק החגור סביב המותניים עשוי למשוך תשומת לב מן הסוכבים. בדברנו למיקרופון אנו נוטים לשנות את צורת התבטאותנו, במודע או שלא במודע. נציגת הפרייקט דיווחה מפי אינפורמנטית אחת בוו הלשון: "הצידור לא מאוד נוח. מסורבל, מושך הרבה תשומת לב ולכן התנהגות הסוכבים מושפעת מכך." ומאת אינפורמנטית שנייה דיווח כך: "המכשיר עורר תגובות אצל אנשים שהייתה במחיצתם. בלא שידעו את נושא המחקר הם החלו לדבר בצורה רכות סבבו בנושא הסביבה והשפעה אחר דרוח כדלקמן: "בעקבות ההקלטה שיחות רכות סבבו בנושא הסביבה והשפעה מנוכחות המכשיר". נשתדל ללמוד מכך. גם לשוננו של האינפורמנט נושא מכשירי ההקלטה עשויה להיות מושפעת מידיעתו את היותו מוקלט. אולם ניסיון שנצבר בכינון מאגרי לשון מדוברת בעולם מוכיח, כי מודעות המשפיעה על הלשון בדרך כלל שוככת אחרי זמן קצר. אולם אנו מבקשים מן האינפורמנטים להיות מודעים לא רק בעת ההולפת הקלטות אלא גם בתוך פרק הזמן שבנייהו, בבקשנו מהם לתעד בקול מי היו בני שיחם ומה הקשר השיח (ר' ציטטה מתוך ההוראות לאינפורמנטים להלן). מניסיוני בהקלטות עצמי אני יכול להעיד, כי מודעותי שלי בעת ההקלטה לא שככה ברוב שעות ההקלטה, שאצלי נמשכה פרק זמן של 48 שעות.<sup>2</sup> שאלות שנשאלתי לגבי טיב המכשירים שאני נושא (אני נשאתי את המיקרופונים תלויים על צידי משקפי לצד האוזניים) העלו תמיהה לאחר תשובות לא מוצלחות במיוחד ("זה וקמן, אבל הוא לא פועל כשאני מדבר איתך" – והרי אני מעולם, מעולם לא השתמשתי בווקמן...). אולם תמיהה זו לא נמשכה זמן רב, והחיים חזרו למסלולם הטבעי ונמשכו כך. התרשמתי מהאונה להקלטות שעשה בני (אף הוא במשך 48 שעות) ומהקלטות של אחרים היא כי אין ניכרת הקפדה יתרה בשימושי הלשון בשל ההקלטה. נושא זה מן הראוי שייבדק במחקר רחב ועמוק. מכל מקום אנו מודגשים בפני האינפורמנטים שעליהם להשתדל להתנתק מן העובדה כי הם מתעדים את שעותיהם, ולהתנהג כרגיל (ר' להלן).

— שאלה נוספת שעלתה שוב ושוב — הן בשלב התכנון והן בשלב גיוס הסוקרים

2 הונונה הראשונית הייתה להקליט פרק של 48 או 72 שעות לכל אינפורמנט. נתברר שאין הדבר מעשי, לא מבחינת העול המוטל על המקליטים, לא מבחינת כלכלית.

של האינפורמנט ומעבירים את הזכויות על ההקלטות לאוניברסיטה, לא נוכל לקבל הסכמתו של כל מוקלט משני להשתמש בהקלטות. הרי כל מקליט יבוא במגע עם אנשים רבים, מהם זרים לו בתכלית, בתוך פרק זמן ההקלטה.

אינפורמנטים בפרייקט דומה נדרשו לחתום את כל בני שיחם מראש על כתב הסכמה. שוו בנפשכם: האינפורמנט עולה לאוטובוס. לפני שהנהג מספיק לומר לו: "יאללה, פנס כבר פנימה, אני סוגר תדלת", יוציא האינפורמנט טופס הסכמה ועט, ויבקש להחתים את הנהג על הטופס. האוטובוס דחוס, והאינפורמנט יידחק עם הסלים אל פנים האוטובוס, ועוד לפני שישמע את הקללה שתופנה אליו, יוציא טופס ויבקש מן המקל: "אולי, בטובך, תואיל לחתום לי על טופס הסכמה לכלול את הקללה שאתה עומד לפלוט עכשיו במחקר של אוניברסיטת תל-אביב?"

התיאור דלעיל מוגזם, כמובן, אולם השאלה של הסכמת בני השיח אינה שאלה פשוטה. (לדבר משאלת האתיקה עומדת גם שאלת המודעות, כפי שנידונה בקצרה כבר לעיל.) מה נעשה אנני האם מותר לנו להשתמש בחומר זה ולפרסמו? החוק הישראלי מקל לעניין זה, ומאפשר הקלטה ושימוש בה בתנאי שאינה הקלטה סתר. הקלטה שאינה בגדר האונת סתר היא כו שניעשתה בדרך שברי כי הדובר יודע שדבריו נשמעים על ידי האינפורמנט, גם אם במשתמע. דהיינו, גם אם האינפורמנט אינו משתתף פעיל בשיחה, אולם שומע את הדברים בשל קרבתו לדובר, ההקלטה מותרת. האם הדבר נכון גם במקרה ראות אתית? זו שאלה פתוחה, אולם מעמ"ד קשוב לשאלה האתית ונקט צעדים כדי להקדוהתה. כבר בתחילת המפגש עם נציג מעמ"ד, נאמר לאינפורמנט ונמסר לו בכתב כדלקמן:

— לאחר איסוף ההקלטות ועריכת הראיון בידי נציגי הפרייקט, יועבר כל החומר בשלמותו לצוות הקמת המאגר, וגישה להקלטות המקוריות ולנתונים תינתן אך ורק לצוות הקמת המאגר. האוניברסיטה שלך ושל כל המוקלטים תישמר על ידי כך שבהקלטות ובתמליליהן לא יוזכרו שמותיכם, כתובותיכם המדוייקות ומספרי הטלפון שלכם, או כל פרט אחר שעשוי להותכם במדוייק. במקום פרטים אלה יופיעו שמות, כתובות ומספרי כרטיים. אני מבקש להודיע ששמירה על פרטיותך ופרטיות קרוביך תברך עומדת בראש מעיינינו. אם בתום ההקלטה תבקש/ מאיננו למחוק קטעים מסויימים שהוקלטו ואינם לרוחך או אף להשמיד את כל החומר שהוקלט — אנו נכבד את בקשתך בלא היסוס.

וכבר נתבקשנו למחוק מן ההקלטה קטע משיחת טלפון אחת. כמובן, קטע השיחה הזה ואף סביבתו נמחקו לפני שהועברה הקלטה לגיבוי. דברים ברוח זו נאמרים ונמסרים בכתב לאינפורמנט גם לאחר ביצוע ההקלטות, ואף בנוגע לראיון הסוציולוגיטי שנערך:

— פרטי הראיון שנערך עמך ייצורו באוניברסיטת תל אביב וישמשו לצורך מחקרי בלבד. חלק מן הפרטים יתועדו בכסיס נתונים שיוצמד להקלטות ולתמליליהן, אולם גם שם לא יכללו שמך, כתובתך המדוייקת ומספרי הטלפון שלך. אני מבקש להודיע ששמירה על פרטיותך עומדת בראש מעיינינו.

ב. הכנת המשתתף לפרק זמן ההקלטות כוללת הסבר מדויק על התהליך, הדרכה טכנית בתפעול המכשירים והדרכה לביצוע ההקלטות ולהתנהגות בזמן ההקלטה. הדרכה זו ניתנת בעל פה, אולם דפי הסבר נשארים בידי האינפורמנט, ובהם כל החומר הדרוש לתפעול ולהתנהגות בעת ההקלטה. למשל:

— אל תוטרד/י אף פעם מהעובדה שהמכשיר עובד והקלטת מקליטה גם אם אין שום דיבור במשך זמן רב. זהו חלק מהחיים הטבעיים שלנו. תן/תני לקלטת לרוץ ואל תפסיק/י או תפעיל/י אותה אלא כשנסתיימה וצריך להחליפה. מלבד בזמני החלפת הקלטת מדי ארבע שעות, פשוט שכח/י את העובדה שאתה מקליטה/ו ורוץ/י עם סדר היום הרגיל בלא הפרעה. זה טוב לך וטוב למחקר.

— במשך כל הזמן שאתה מקליטה/ה אנא תערי/י היכן אתה נמצא/ית ומה תנועותיך. אמור/אמרי אל תוך מכשיר ההקלטה בקצרה: עכשיו אני בכית, ברחוב אלנבי, בדרך לסבתא, באוטובוס, בדרך לשוק הכרמל, בשוק מחנה יהודה, באוטו של אחותי, בסופרמרקט, בחנות בגדים ברחוב סוקולוב, נכנסת לשיעור במכללת ספיר.

— אם אפשר, אמור/אמרי — בצורה דיסקרטית בהיותך לבד או כשלא שומעים אותך — עם מי ריבית או עם מי אתה הולך/כת לדבר. אם אתה יודע מה רקע האדם שעמו שוחחת או שאת דבריו שמעת (כלומר: מוצא, גיל והשכלה), אמור/אמרי גם זאת. נתונים אלה עשויים לעזור לנו במחקר, אולם אינם הכרחיים. אמור/אמרי אל תוך מכשיר ההקלטה: אני נכנסת לדבר עם המורה של ערן בני, שהיא באמצע שנות ולומד באוניברסיטה; אני שהיא ממוצא תימני; שוחחתי עם הנהג באוטובוס שהיה בן החמישים שלה ונראה לי שהיא ממוצא תימני; שוחחתי עם הנהג באוטובוס שהיה בן 40 בערך ונראה לי שהוא ממוצא רוסי. אם שוחחת עם אותו אדם כמה פעמים במשך הזמן שבו אתה מקליטה/ה, אין צורך לחזור על הנתונים הנוספים מדי פעם, רק בפעם הראשונה. איזכור שם האדם יספיק כפעמים הבאות. אל דאגה: שמוותיהם ופרטים מזהים של כל המוקלטים ישונו בהקלטה ובתמליל לפני פרסומם.

— שים/שימי לב: יש לנו נטייה ללחוץ או לדבר בשקט רב אל המיקרופון כשאנו לבדנו או כשאנו חוששים שמישהו ישמע. מניסיונו, דיבור שקט מדי לא נקלט כהלכה, ולכן אנא דבר/י בקול לא שקט מדי גם כשאח/ה לכה. אם יש לך חשש שמישהו ישמע וישאל שאלות מיותרות — תורי/י על הקלטת הנתונים האלה לעת עתה או ככל.

בכל מקרה, ככל שכיבה, ככל מפגש ועם כל אחד עדיף לנהוג כרגיל ולהיראות כמי שמתנהג כרגיל. אין צורך להחביא ולהסתיר במיוחד את השתתפותך במחקר, אולם מטרננו לתעד מהלך חיים טבעי ושוטף, ולא כזה שמסתובב סביב המחקר. (ההדגשה במקור)

נציג מעמ"ד ייצא את בית האינפורמנט כשמכשיר ההקלטה פועל, וישאיר בידי האינפורמנט מספרי טלפון לעזרה. הוא ישוב ויודא בטלפון אחרי תום קלטת אחת (4 שעות) שהקלטת הוחלפה וכי לא התעוררו בעיות בדרך. אם נתגלו קשיים, ידאג נציג הפרוייקט לעזור לאינפורמנט להתגבר עליהם, בין אם בטלפון, בין אם בכיקור חוזר.

ג. פרקי הזמן להקלטות שיועדו לאינפורמנטים נעים בין שמונה לעשרים וארבע שעות רצופות. מכשירי ההקלטה המשמשים להקלטות הבריחה הטרונית הם הטובים מסוגם. אלו הם מכשירים דיגיטליים TCD-D100 (DAT) מתוצרת Sony. המיקרופונים המשמשים אותנו בהקלטות הם מיקרופונים סטריאו דיגיטליים DSM-IS/L מתוצרת Sonic Studios, המיועדים לגשייה קרוב לאוזניים, כך שכל מה ששומע האינפורמנט מוקלט בצורה זוהה. איכות ההקלטה של המיכשור הזה יוצאת מן הכלל, אולם הרכבתו ותפעולו דורשים מיומנות מסוימת. האינפורמנט נושא את הציוד בתיק ייעודי שנחגך סביב למוטנניו, ומעבידי את כבל המיקרופון מתחת לבגדיו כך שהמיקרופונים, שייצבנו אותם על קשת נשיאה התלויה על עורפו, מצויים בקרבת אוזניו. אנו משתמשים בקלטות דיגיטליות Sony 120P שמשך ההקלטה בהן ארבע שעות (קלטות דיגיטליות מוקלטות בצד אחד בלבד). כל קלטת תסומן כסימון שרירותי של האינפורמנט (למשל: במיספור רציף), במספר הקלטת על פי סדר, ובתאריך ושעות ההקלטה, שאותם יוסיף האינפורמנט.

בתוך פרק זמן ההקלטה, החלפת קלטת היא החליה החלישה בתפעול, ולפיכך החלטנו לברוק את התהליך ראשית עם החלפת קלטת פעם אחת בלבד. ארבעת האינפורמנטים הראשונים מתוך השישה עשר שבדוק כל מכוני יקליטו את אורח חייהם. אם כן, במשך שמונה שעות רצופות. ארבעת האינפורמנטים הבאים יקליטו במשך שתיים עשרה שעות, דהיינו יבצעו שתי החלפות קלטת. אחרי צבירת הניסיון הזה, שאר שמונת האינפורמנטים יתבקשו להקליט את אורח חייהם במשך 24 שעות. יממה זו כוללת את שעות השינה של האינפורמנט, כך שסביר שבפועל יתקבלו לא יותר מאשר 16 שעות הקלטה בארבע קלטות. הקלטה ארוכה מפרק זמן זה אינה ישימה.

ד. בתום פרק זמן ההקלטה שיועד לאותו אינפורמנט, ישוב ויבקר נציג מעמ"ד בביתו. הוא יודה לו על השתתפותו ויתשאל אותו בדבר התנאים והנסיבות של ההקלטות שנעשו: באיזו סביבה נעשתה כל הקלטה ומי היו בני השיחה שהשתתפו בה? מה הקשר של בן השיחה אל המקליט? מה שמו (למען הזיהוי בהקלטה; אנו נחליף בשם בדוי), מוצאו (ארץ לידה, עדה, גיל, מין, ומצב משפחתי), והשכלתו. אם לא יודעים פרטים אלה, יתבקשו האינפורמנטים לספק השערות או קירובים.

ה. אחרי כן יערוך הסוקר ראיון סוציולוגי עם האינפורמנט. בראיון זה יבדק הרקע חברתי והלשוני של האינפורמנט: מקום מגורים נוכחי, מקום מגורים קודם ומקום המגורים בילדות; מוצאו ומוצא הוריו והורי הוריו, ושנת עלייה לארץ אם אינו יליד הארץ; לאום ועדה; דת ומידת הקשר לדת; השכלה, תעסוקה ותעסוקת ההורים בשנות ילדותו; מצב כלכלי; שירות צבאי; עמדה פוליטית; שהייה ארוכה בחו"ל; לשון הדיבור העיקרית ולשון הדיבור בכיתה; לשון דיבורם של ההורים; לשון הדיבור עם בני המשפחה הקרובה בילדות; רמת העברית בקריאה, בכתיבה ובדיבור; וידיעת שפות אחרות.

השאלות על ידיעת העברית נוספו כדי לברוק הערכת הדוברים את העברית שום מדברים. יחס הדוברים לשפה צופן בחובו השלכות משמעותיות לכיווני התפתחות השפה והשתנותה ולמערכת החברתית והתרבותית ככלל (ר' מחקרו הקלאסי של

1963 Labov; וכן, למשל, 1987 Fasold פרק 6; 1996 Preston). מעמ"ד אינו המקום למחקר ייעודי לשאלות אלה. במסגרת זו רצינו לברוק רק את הערכתם של דוברים – בעיקר ילידיים – את רמת דיבורם בעברית. השערת מחקר ראיה לבריקה היא כי הערכת דוברי העברית הילידיים את לשונם נמוכה יחסית. כמה תשובות שנתקבלו מלמדות כי יש מקום לבריקה הנושא. השוואה לרמת העברית הניבטת מן ההקלטות להערכת הדוברים את לשונם עשויה להיות מעניינת ובעלת השלכות ניכרות, במיוחד לדרך החינוך הוראת העברית בבתי הספר.

הנתונים שיתקבלו מן התשאל הסוציולינגוויסטי הזה ייכל במעמ"ד ויעמדו לרשות משתמשי במסד הנתונים הצמוד להקלטות. כפי שנאמר לעיל, האינפורמנטים ובני שיחם יזוהו בשמות בדויים ופרטיים תישמר. כך בהקלטות ובתמליליהן, כך במסד הנתונים הסוציולינגוויסטי. כמובן, אפשר יהיה לזהות את הדוברים על פי קולם, אולם הסיכוי לפגישה בין משתמשי המאגר לבין האינפורמנטים שהתנדבו לפרויקט קלוש ממילא. בכל מקרה, לאינפורמנט הזכות לחוק קטעים גם בדיעבד, כך שמה שיוותר לא יהא בו מבוכה לדובריו.

1. במסגרת הבריקה הטרומוית הוספנו עוד שאלות שנועדו לברוק את רכיב העבודה: איך הסתדרו האינפורמנטים עם נשיאת המכשירים, איך הסתדרו עם תפעולם, האם הבינו את הכתוב בספסים שנמסרו, והאם הבינו את כל השאלות בראיון. הנשאלים מתבקשים גם להעיר על הקשיים ולתרום לנו כך מניסיונם.

2. עם תום התשאל, חותם האינפורמנט על כתב הסכמה בנוסח זה:

לאחר שהובאה לידיעתי זכותי לשמוע את ההקלטות /או לקרוא את תמליליהן לפני שהן מיוספות למאגר העברית המדוברת בישראל (מעמ"ד), אני מביעה בואת את הסכמתי שההקלטות שמסרתי ישמשו, כולן או חלקן, כחלק ממאגר העברית המדוברת בישראל (מעמ"ד) של אוניברסיטת תל-אביב.

עוד אני מביעה בואת את הסכמתי שפרטי הראיון שנועד עמי יועדו בבסיס הנתונים של מאגר העברית המדוברת בישראל (מעמ"ד), ובלבד ששמי, כתובתי, מספרי הטלפון שלי ופרטים מזהים אחרים לא יוכנסו למאגר הנתונים, כך שהגישה לנתונים אלה תוגבל לעוסקים בפועל בהקמת המאגר.

הסכמתי ניתנת בכפוף לכך שמטרתו של מאגר העברית המדוברת בישראל (מעמ"ד) היא מחקרית, ותפוצת המאגר ופרסומיו – בכל אמצעי שהוא – ייעשה ללא מטרת רווח. אני מאשרת בואת כי בתמורה להשתתפותי קיבלתי מאת אוניברסיטת תל-אביב כאמצעות נציג/ת הפרוייקט סך של – ### ש"ח.

סכום הכסף הניתן לאינפורמנטים ניתן כתגמול על השתתפותו בפרוייקט, ואף משמש כתהליך של מתן הרשות לצוות מעמ"ד להשתמש בחומר המוקלט. בהודמנות זו אציין, כי התגמול הכספי משמש אף הוא, כפי שכבר נזכרנו, כתמריץ למתבקש לשמש כאינפורמנט, במיוחד עבור צעירים ומעוטי אמצעים. וכבר ראינו לעיל שהיו מן האינפורמנטים שהציעו לבטל את מתן התגמול אם חשו שהתמורה שתיוסר אינה

מספקת. מעמ"ד מתגמל את המשתתפים בכל מקרה, גם אם לא ימצא בהקלטותיהם חומר ראוי להכללה במאגר.

ח. הקלטות החומר הכתוב מובאים עתה אל משרדינו. כל החומר, המוקלט והכתוב, נרשם במסד נתונים ומוגובה. בשלב זה גיבוי ההקלטות נעשה על ידי העתקת החומר המוקלט אל הקלטורים.

ט. התקליטורים נסקרים לבחירת החומר לתמלול. כדי להבטיח ייצוג אמין של השונות הלשונית, הדמוגרפית ותלוית הקשר כאחת, היה טוב לו אפשר היה לשמור את עקרון האקראיות בעת דיגמת כל הנתונים הסקסטואליים כדלקמן: לאחר איסוף כל ההקלטות של האוכלוסייה הנרגמת דיגמה דמוגרפית תבצע דיגמת הקשרי השית. פרקי הזמן בן 24 השעות יתחלקו שווה בשווה בין כל האינפורמנטים. אידיאלית, יוקלטו שבעה רציפים שווים כאלה, כשכל רצף פותח ביום אחר בשבעה. מכל הקלטה בת יממה ייבחר באקראי קטע של טקסט באורך של שעה, והוא שיהווה מקור לנתונים שייכללו במאגר הראשי. בחירה אקראית זו עשויה לשקף את כיוור מצבי השחי הטבעיים, הן על פי היקרויותיהם בזמנים שונים, הן בהתאם להקשרי השחי, והן בהתאם לסוגי השחי המזומנים לסוגי אוכלוסייה שונים.

מניחות הקלטות שכבר נעשו נראה כי דיגמה אקראית של הזמן תהיה קשה, אם לא בלתי אפשרית, אם מטרתנו להשיג תאים (דהיינו: יחידות בנות 5000 מילה) שיהיו אחידים לשם מחקר סוגי הלשון. המגבלות הן טכניות, סטטיסטיות וסוציולינגוויסטיות, והן בעיות המשולבות זו בזו. בעיה ראשונה היא תדירותם, צפיפותם ואיכותם של קטעי דיבור שאפשר לכוללם במאגר. יחידות בנות חצי שעה או שעה שיידגמו אקראית סיכוייהן קטנים להכיל קטעי דיבור אחידים ונוחים לתמלול. לא אחת מן הקלטות שברשותנו מכילה – בפרק זמן של ארבע(1) שעות – רק מוזיקה מן הרדיו או שדרורי טלוויזיה. במקרים אחרים רעשי רקע גורמים לקולו של האינפורמנט ובני שידורו להיבלע בהם. גם שיחה שבה חפיפות הרבה, ווראי שיחה בת יותר משניים או שלשה אנשים, בעיקר בתרבות הישראלית, היא בלתי אפשרית לתמלול בדרך כלל. נצטרך, אם כן, לפני דיגמת הזמן, לדאוג לסריקת הקלטות ולהסרת קטעי שתיקה וקטעי דיבור שלא ניתן לפיענח. לפיכך, פרק הזמן שיידגם לא יוכל להידגם על פי הרצוי באקראי מתוך יממה, וייבחר לאחר ניפוי.

תדירות הדיבור של האינפורמנט בתוך קטעי שיח אף היא מהווה שיקול בבחירת הקטעים להכללה במאגר. גם בהקלטותיו של אינפורמנט שהוא מרצה במקצועו ואשר משום כך יוכל לספק פרק זמן ברה-הכללה שבו מונולוג בנושא מסוים, או שיחות עם תלמידים, גם שם דיגמה אקראית של פרק זמן לא תביא בהכרח את פרק הזמן הראוי לקבלת תא ראוי מאותו סוג. אינפורמנט שהוא טלפוני בבני יספק שיחות יקלות כמעט מונוטוניות וחסרות ערך ממש לניתוח לשוני, מה עוד שמכשיר ההקלטה יקלוט כמעט אך ורק את דיבורו שלו: ניסיונו מלמד שאם בני השיח שמעבר לקו אין שומעים, בדרך כלל (אנו נצטרך, כמובן, לתת את דעתנו כיצד אפשר לפתור שאלה זו, מן הצחינה השנית או בררכים אחרות. כמובן, אי אפשר למעמ"ד שלא יכלול שיחות טלפון).

דוגמה אחרונה לעניין זה היא פסיכולוגית, שדגימה אקראית תספק לנו מונולוגים של לקוחותיה ולא את דיבורה שלה, והלוא הפסיכולוגית היא שעלתה במידגם האוכלוסין לכינון המאגר, ולא לקוחותיה. לפיכך נראה לי כי נזדקק לא רק לניפוי הקלטות אלא אף לברירה ובחירה של ממש של פרקי הזמן שייכללו במאגר. לקראת כינון המאגר כולו נבדוק אפשרות לגיבוי וסריקה מתוחכמים יותר, כך שלא נצטרך להקדיש זמן יקר ומיותר להאזנה לחומר שאינו מתאים למאגר, במיוחד בפרקי זמן שאין בהם דיבור כלל.

יא. לאחר שנבחר פרק הזמן להכללה במאגר, יירשם במסד נתונים שבו יפורטו נסיבות ההקלטה, מקומה, הרוברים בה, ומידע נוסף כדבר הקשרי השיח המתועדים בפרק הזמן הזה. שלב זה עוד לא הגיע לידי מימוש כעת כתיבת דברים אלה.

יב-טו. תמלול ההקלטות הוא שיעור אחר לגמרי, ופוחה שערי עיון רחבים ועמוקים. בעל היובל יוכל להחכימו רבות בשאלות של תמלול ותעתיקים. על כך, ועל שאר הלוקחים שיופקו מן הבריקה הטרומית, נביא בפני הקוראים כאכסניית אחרת דבר דבר בעיתו.

#### כיבולוגרפיה

יזעאל, ש', ב' הרי "לקראת כינון מאגר העברית המדוברת בישראל", לשוננו ס"ד, וג' רהב (תשס"ב) עמ' 265-287.

Fasold, R. 1984 *The Sociolinguistics of Society. Introduction to Sociolinguistics Volume 1.* (Language and Society, 5.) Oxford.

Izre'el, Sh., B. Hary and G. Rahav 2001 "Designing CoSIH: The Corpus of Spoken Israeli Hebrew". *International Journal of Corpus Linguistics* 6/2, pp. 171-197.

Labov, W. 1963 "The Social Motivation of a Sound Change". *Word* 19: 273-309. = *Sociolinguistic Patterns.* (Conduct and Communication No. 4). Philadelphia 1972, pp. 1-42.

Milroy, L. 1987 *Observing and Analysing Natural Language: A Critical Account of Sociolinguistic method.* (Language in Society, 12.) Oxford.

Preston, D., 1966 "Whaddayaknow". *Language Awareness* 5/1, pp. 40-47.

דף האינטרנט של מעמ"ד:

<http://spinoza.tau.ac.il/humanities/semitic/maamad.html>

אתר האינטרנט של המאגר הלאומי הכריטי: <http://info.ox.ac.uk/bnc/>

מכנה תת-המאגר של האנגלית המדוברת מתואר כך: [http://info/ox/ac/uk/bnc/what/spok\\_design.html](http://info/ox/ac/uk/bnc/what/spok_design.html)