

שלמה יזרעאל, בנימין הרי וג'ורא רהב

לקראת כינון מאגר העברית המדוברת בישראל

1. מבוא

1.1 הגישה הבלשנית המבוססת על חקר מאגרי לשון (corpus linguistics) זכתה לתשומת לב רבה במהלך עשרים השנה האחרונות. בזכות גישה מחקרית זו תיאור הלשון והסקת המסקנות התאורטיות מתבססים על אמצעים סטטיסטיים ועל בדיקת שימושי הלשון הבאים לידי ביטוי במציאות. צורה בסיסית מאוד של שיטות מחקר שהתבססו על מאגרים טקסטואליים כבר נודעה בעבר; ענפי החקר המגוונים של הלשונות השמיות, למשל, התבססו בעיקר על טקסטים כתובים. מקאנרי ווילסון (1996, עמ' 2–4) מזכירים כמה עבודות בתחום הבלשנות הכללית העשויות להיחשב ניצניה של הבלשנות מבוססת המאגר: חקר לשונם של ילדים מתוך יומנים כתובים שערכו הוריהם בשלהי המאה התשע-עשרה ובתחילת המאה העשרים; מחקרי שדה בשיטתו של בואז (1940) או בשיטותיהם של בלשנים מאוחרים יותר מהאסכולה הסטרוקטורליסטית. עם גבור השפעתן של התאוריות מבית מדרשו של

* עיקרי הדברים שבמאמר זה נישאו מפי שלמה יזרעאל בהרצאה בפני מליאת האקדמיה ללשון העברית ביום כ"א בכסלו תשס"א (18 בדצמבר 2000). אנו מודים לפרופ' משה בר-אשר, נשיא האקדמיה ללשון העברית ועורך "לשוננו", על פרסום המאמר.

המאמר נכתב בסיועם של ג'ון די בואה ושל מירה אריאל, ותרומתם היא בעלת משמעות בעיקר לכינון תבנית נסיבות השיח. תודה להם ולשאר חברי צוות היועצים של מאגר העברית המדוברת בישראל על הערותיהם החשובות. אנו מודים גם למרגלית מנדלסון שעזרה בהרקה ראשונית של דפים אלה אל כלים הטבעי. כן אסירי תודה אנו לרגינה נרום, שסייעה לנו בהבהרת כמה היבטים סוציולוגיים (ראה נרום, ברפוס). יזרעאל והרי הרצו על מאגר העברית המדוברת בישראל (מעמ"ד) ותכננו במסגרות שונות ושמעו הערות בונות והצעות מאנשים רבים. תודה מיוחדת לאלנה טוניני-בונלי ולג'ון סינקלר מ-Tuscan Word Centre על מה שלמדנו מהם ועל עידודם המתמיד.

צוות החוקרים של מאגר העברית המדוברת בישראל: שלמה יזרעאל (ראש התכנית וחוקר ראשי); בנימין הרי (חוקר ראשי); ג'ון די בואה (אנליסט המאגר); מירה אריאל (חקר השיח ופרגמטיקה); ג'ורא רהב (סוציולוגיה וסטטיסטיקה). צוות יועצים: אליעזר בן רפאל (סוציולוגיה ולינגוויסטיקה – היבטים סוציולוגיים); יעקב בן-טולילה (סוציולוגיה ולינגוויסטיקה – היבטים בלשניים); אוטו יסטרוב (תיעתוק, פונולוגיה ודיאלקטולוגיה); שמואל בולוצקי (פונולוגיה, מורפולוגיה); ג'פרי כאן (תחביר); אילנה שוהמי (חינוך לשוני).

מאמר ההמשך המתפרסם בחוברת זו (ש' יזרעאל, להלן, עמ' 289–314) מביא דוגמאות טקסטיות מן המאגר.

נועם חומסקי קטן המשקל הסגולי של ההתבוננות האמפירית, וחקר מבנה הלשון התמקד בהעמקה סובייקטיבית של החוקר בלשונו הוא ולא דווקא באיסוף נתונים אובייקטיביים. הגישות שהתפתחו בהשראת תורתו של חומסקי, הנחשבות מובילות בתחום חקר הלשון עד ימינו אנו, אינן מתמודדות עם מגוון רחב של נושאים, ובייחוד לא עם השונות הלשונית. שונות זו מתבטאת בקיומם של דיאלקטים גאוגרפיים, חברתיים ואתניים, בחלופות (=וריאנטים) בשפה מדוברת מול אלה שבשפה כתובה ובמשלבי הלשון של בני המינים השונים, של בני רבדים סוציו-אקונומיים שונים, גילאים, מקצועות וכיוצא באלה. כיום השונות הלשונית נחשבת הן בבלשנות העיונית והן בבלשנות התיאורית כמאפיין מהותי, אורגני, ללשון האנושית.

המשימה הראשונה שייטול על עצמו הבלשן היא כמובן איסוף טקסטים, כינון מאגר לשון. ההתפתחויות האחרונות במדעי המחשב, השינוי המכריע באפשרויות האחסון וגידול הזיכרון במחשב שיפרו במידה ניכרת את תנאי חקר מאגרי הלשון בעולם כולו. כיום נתפס מאגר הלשון כמאגר ממוחשב, כאוסף של טקסטים ממוחשבים. מאגר לשון טוב הוא מאגר המכיל טקסטים ממוחשבים שנדגמו כדי לתת ייצוג מרבי של שפה או של ניב.

1.2 בעוד מאגרי לשונות שונים ורבים הולכים ונאספים בעולם כולו¹ עדיין אין בנמצא מאגר של העברית בת זמננו. יתר על כן, חקר העברית החדשה – ובעיקר חקר משלביה המדוברים – לוקה בחסר בתחום התיאור הבלשני, וזאת בין השאר בשל מחסור בנתונים (קדרי, תשמ"ד). איש לא יחלוק על הצורך במחקר אמפירי שיעמיד את ידע הלשון העברית בת זמננו על יסודות מוצקים. אולם כבר חלפו מאה שנות שימוש בעברית, ומחקר העברית עודנו חסר. תפקודה המחודש של העברית כשפת היום-יום העצים זרימה בלתי פוסקת של פרסומים: מילונים, מחקרי דקדוק, ספרי לימוד ועוד. רובם כווננו להוראת השפה, ומעטים מאוד כיום המחקרים הלשוניים והדקדוקיים המבוססים על השימוש בה. מחקר שישקור את ההיסטוריוגרפיה הלשונית של העברית החדשה עשוי להיות בעל עניין רב. מכל מקום, המחקרים הקיימים, בניגוד למחקרים שעסקו בנדבכים קודמים של העברית, לא הושתתו על מאגרים מקיפים. מאגר מידע הוא תנאי הכרחי לתכניות אחרות, שבלעדיו אינן בנות השגה, יהא זה דקדוק של העברית החדשה, מילון מקיף, או כל תכנית מחקר או פיתוח אחרת לצורך עיוני או לצורך מעשי. אפשרויות המחקר הגלומות במאגר שיטתי של שפה הן עצומות וכוללות יישומים בתחום הבלשנות, התרבות והחברה, וכן יישומים טכנולוגיים רבי עצמה. צורך זה הוא שהביאנו להתמודד עם המשימה של כינון מאגר מקיף של העברית הישראלית.

1. אדוארדס, 1993; < <http://www.ruf.rice.edu/~barlow/corpus.html> > .

קריאות לאיסוף מאגר מדעי של העברית המדוברת החלו להישמע זה זמן מה. בן-טולילה (תשמ"ט) תיאר את מאגר הצרפתית במונטריאל (Le corpus Sankoff- Cedergren du français parlé à Montreal) וקרא לכינון מאגר של העברית המדוברת. ברשימת הביקורת שלו על ספר דקדוק העברית של גלינרט (1989), דקדוק המבוסס על עדותם של שישה אינפורמנטים, קרא בלאו (תשנ"א) להשתתף דקדוק מקיף על מאגר נרחב של חלופות כתובות ומדוברות של השפה ולא על קביעות המבוססות על כשירות לשונית מפי דוברים ילידים ספורים. חשוב לציין שבדיקה אינטרוספקטיבית כגון זו שהדקדוק של גלינרט מבוסס עליה (כשהכוונה היא להציג כשירות לשונית יותר מאשר תיעוד של ביצוע) אינה יכולה לשמש בסיס לניתוח מקיף אמתי של שפה, ניתוח שיכלול את כל רצף החלופות הלשוניות שלה.

לפני עשרים שנה ויותר הגיש מ"צ קדרי הצעה לוועדת האקדמיה לכינון מאגר של העברית בת זמננו בהציעו להתמקד בסקר לשון הספרות (בלאו, תשנ"ו, עמ' 11). בדבריו בקונגרס העולמי התשיעי למדעי היהדות ראה קדרי חשיבות בכינון מאגר של העברית הכתובה והמדוברת כאחת בפרטו את חשיבותה של "פונותיקה ללשון הדיבור המוקלטת" (קדרי, תשמ"ח, עמ' 91). בהרצאתו לכבוד שנת הלשון העברית חזר קדרי על קריאתו לכינון מאגר של העברית החדשה, אולם כאן הוא ראה דחיפות יתר דווקא בכינון מאגר של העברית הספרותית החיה (קדרי, תשנ"ו). בין נימוקיו הדגיש קדרי בעיקר את הצורך ביצירת תשתית למחקר. נקודת זינוק טובה לעבר המטרה הזאת היא עבודתו של יעקב שויקה מאוניברסיטת בראילן, השוקד כעת על איסוף מאגר ממוחשב של העברית הכתובה. מאגר בראילן לעברית מודרנית (מב"ע) כלל בסוף שנת 1999 עשרים ושישה מיליון מילה.² העברית הספרותית, ועמה חלופות נוספות של השפה הכתובה, הן בבחינת קיים הניתן לאיסוף בכל עת, לכן הענקנו עדיפות ראשונה לתיעוד העברית המדוברת.

2. מטרות

א. יצירת מאגר של העברית הישראלית המדוברת כתשתית למחקר שיטתי. מחקר המאגר יקיף מגוון רחב של נושאים הקשורים בשפה העברית ובמתודולוגיה הכללית של חקר הלשון המתבסס על מאגרי לשון.
 ב. הפצת המאגר לציבור במולטימדיה ובדפוס. ההפצה באמצעים אלקטרוניים – תקליטורי שמע, תקליטורי די-וי-די והאינטרנט – תיעשה כך שהקלטות ותמליליהן יוצגו במקביל ובשילוב דרכי תיעוד וניתוח נוספות.

2. מאגר זה אינו נגיש לעת עתה לקהיליית החוקרים ואינו מתיימר להיות מייצג. אנו מודים ליעקב שויקה על הנתונים שמסר לנו.

ברצוננו להדגיש שמאגר העברית המדוברת בישראל (מעמ"ד) יהיה נגיש ופתוח לכול ולכל צורך.

3. מהות המאגר

3.1 מאגר העברית המדוברת בישראל (מעמ"ד) ייצג את מגוון חלופות הלשון של העברית המדוברת בישראל היום. בכוונתנו לכלול מדגם מייצג של חלופות דמוגרפיות ושל חלופות נסיבתיות. החלופות הדמוגרפיות מזהות עם קבוצות דוברים שונות: גאוגרפיות, אתניות, סוציו-אקונומיות וחברתיות (גיל, מין, השכלה, מקצוע, נטייה מינית וכו'). חלופות נסיבתיות הן פונקצייה של מצבים שונים, כגון שיחה (פנים-אל-פנים, טלפונית), סוגי אינטראקציה (היחסים הבין-אישיים הכרוכים בה, מבנה השיח הרלוונטי) ונושאי השיח וסוגיו (ספונטני, מתוכנן, מצוטט מן הכתב).

ככיוון המאגר הבאנו בחשבון את המבנה המיוחד של החברה הישראלית, שבה 80.1% יהודים ו-19.9% לא-יהודים (14.6% מוסלמים; 2.1% נוצרים; 3.2% אחרים, על פי אומדן מ-1996). 62% מבין היהודים הם ילידי הארץ (26% נולדו לאב יליד הארץ, 22% נולדו לאב יליד אסיה או אפריקה ו-14% נולדו לאב יליד אירופה או אמריקה)³, 12% הם ילידי אסיה או אפריקה ו-25.6% הם ילידי אירופה ואמריקה.

ובכן, התפלגות הדוברים הילידים והלא-ילידים בקרב האוכלוסיה הדוברת עברית בישראל אינה שכיחה. בקרב היהודים 61% הם דוברי עברית ילידים.⁴ בקרב האוכלוסיה הדוברת עברית מספר הדוברים הילידים והלא-ילידים שווה. עובדה זאת, יחד עם המורכבות שנוצרה עקב ההיסטוריה המיוחדת של העברית החדשה, הן בעלות חשיבות מכרעת לגבי הרכב הטקסטים שייכללו במאגר. החלטנו לכלול טקסטים של דוברים ילידים ושל דוברים לא-ילידים כאחד. בישראל אישים בעלי השפעה שאינם נמנים עם הדוברים הילידים (כגון הסופרים אהרון אפלפלד וסמי מיכאל, ראשי הממשלה לשעבר שמעון פרס ויצחק שמיר, הרב עובדיה יוסף, חברי כנסת ורבים אחרים). זאת ועוד, האוכלוסיה הולכת ומתרחבת תדיר בעקבות הזרימה המתמדת של עולים אל ישראל. אף זה גורם בעל השפעה על תנודות במערכת הלשונית שחשוב לתעדו. גם לערבים אזרחי ישראל חלק רב בשימוש בעברית.⁵ ברור

3. ראה הערה 12 להלן.

4. הפקת הלשון של אנשים שעלו לארץ בשנים המכריעות לרכישת השפה עשויה להיות זהה לזו של דוברים ילידים. בהבחנה זו אין משום הבעת דעה על "איכות" השימוש בשפה או על עושר השפה, שכן דוברים לא-ילידים עשויים לרכוש מיומנות מפליגה בשפה הנרכשת.

5. ראה לדוגמה את השפעתו החשובה של "ערבסקות" לאנטון שמאס או את העברית שבפי חברי כנסת ערבים מסוימים.

אפוא כי תיעוד לשונם של הדוברים הילידים בלבד לא ישקף את העברית בת זמננו, וודאי שלא ישקף את המצב החברתי-לשוני בישראל על כל מורכבותו. אילו התעלמנו מהדוברים הלא-ילידים, היה המאגר מעוות מבחינת ייצוג המערכת הלשונית בישראל, והיה פוגם באפשרות לבסס מחקר חברתי-לשוני כלשהו על נתונים בני משמעות.

מעמ"ד ישקף מצב סינכרוני של העברית המדוברת בישראל. היות שמדובר בשפה טבעית, ובייחוד בעברית הישראלית המשתנה בתכיפות, יש להשלים את מעמ"ד בפרק זמן קצר ככל האפשר. כפי שנזכר לעיל ויפורט עוד להלן, שאיפתנו היא ליצור מאגר שייצג נאמנה את כלל סוגי הדיבור העברי בישראל ויכלול מגוון רחב ככל האפשר של דוברים ושל נסיבות שיח. השם שניתן למפעל, "מאגר העברית המדוברת בישראל", משקף שאיפה זו.⁶

3.2 ממדי המאגר

מאגרי הלשון הממוחשבים העומדים לרשות החוקרים בעולם שונים בגודלם. כאשר מאגר כולל טקסטים כתובים ומדוברים, חלקם של הטקסטים הכתובים עולה על שני שלישים מכלל הטקסטים בו. ודאי שיש בכך עיוות התפוצה הסטטיסטית של שני סוגי טקסטים אלה במציאות. הלוא אפילו בעם הספר הלשון המדוברת נפוצה הרבה יותר, והיא בשימוש נרחב הרבה יותר מאשר הלשון הכתובה. עיוות זה מובן מאליו, ואולי אף בלתי נמנע, משום הקלות היחסית שבאיסוף טקסטים כתובים. מעמ"ד יכלול טקסטים מדוברים בלבד, ושאיפתנו היא ליצור מאגר גדול יחסית, כדי לייצג נאמנה את המצב הלשוני של העברית בישראל כיום ולאפשר כך מחקרים בלשוניים וחברתיים מגוונים. מטרתנו היא לכוון מאגר של חמישה מיליון מילה. מעמ"ד יורכב מאלף תאים או רצפים מוקלטים (ראה § 5.1 להלן); בכל תא יהיו חמשת אלפים מילה (כשלושים דקות של דיבור רציף). רצף של חמשת אלפים מילה נראה ארוך דיו כדי לאפשר הסקת מסקנות בלשוניות מהימנות.⁷ 5% מהתאים יוקלטו בוודא. כפי שיוסבר בהמשך, מספר התאים ומספר המילים שייכללו בכל אחד מהתאים ייקבעו מתוך התאמה לייצוגם האמתי באוכלוסיה.

6. תואר השם מדוברת משמש עבורנו כתואר בלתי מסומן לכל סוג של דיבור פָּה, בין טבעי, בין ספונטני ובין שהוא קרוי מן הכתב, וכל מה שביניהם. במילים אחרות, הצירוף לשון מדוברת משמש לסימון ערוץ התקשורת שבדיבור על כל חלופותיו, ועומד לעומת ערוץ הלשון הכתובה על כל חלופותיו. את שם התואר דבורה אנו מייצגים למרחב חלופות השפה הטבעית והספונטנית.

7. בהתבסס על כימותים של מאפיינים לשוניים שנעשו על דגימות טקסטים בני אלף מילה מתוך שלושה מן המאגרים המוקדמים של האנגלית, הסיק בייבר (1990, עמ' 261) ש"הטקסטים בני אלפיים ובני חמשת אלפים מילה במאגרים הסטנדרטיים הנם ייצוגיים ומהימנים דיים על פי כל קנה מידה לניתוח מן הסוג הזה".

מאגר של חמישה מיליון מילה בשפה המדוברת גדול דיו כדי לשקף הן את המבנה הכללי והן את המאפיינים הייחודיים של רוב החלופות הלשוניות. רבים מהמאגרים הקיימים כוללים הרבה פחות מחמישה מיליון מילה ואינם מהווים מאגר לשוני מייצג. מאגרים גדולים יותר התייחסו לשאלת הייצוג רק באופן חלקי.⁸ כינון מאגר של שפה מדוברת שיכלול יותר מחמישה מיליון מילה אינו מעשי. בעוד שטקסטים בשפה כתובה ניתן לאסוף בקלות יחסית בשימוש בטכניקות של סריקה או בשימוש בחומרי אינטרנט, הנה טקסטים מדוברים קשים הרבה יותר לאיתור, להגדרה ולהפיקתם לנגישים לשימוש בשל המורכבות הרבה שבליקוטם ובתמלולם.

3.3 אפשר להציג מאגר לשוני גדול בצורות שונות: בהקלטות מתומללות, בתעתיק פונטי, בצירוף גלוסות ובתרגום. ניתן להוסיף לו תיוגים שונים: לפי מיון תחבירי, מורפו-סינטקטי, מורפולוגי או מורפו-פונולוגי. ניתן להקליטו בקלטות שמע ובקלטות וידאו. לכל דרך רישום יתרונות וחסרונות משלה. לדוגמה, יתרון ברור של הקלטת וידאו הוא יכולתה להציג מאפיינים חוץ-לשוניים, אך יש בה גם מהחיסרון, שהרי האינפורמנטים המיוצגים בה מוגבלים ביכולתם להיות טבעיים כדיבור בגלל טיבו של האמצעי הזה דווקא.

מעמ"ד יכלול את המאפיינים דלקמן:

- הקלטות קול דיגיטליות
- מבחר הקלטות דיגיטליות בוידאו
- תמלילי כל הטקסטים שבמאגר
- תעתיק פונטי של קטעים נבחרים
- ניתוח מורפולוגי רציף של קטעים נבחרים
- תרגום לאנגלית של קטעים נבחרים

4. המדגם המייצג

4.1 על מאגר לשון מייצג להציג שני סוגי חלופות לשוניות: חלופות דמוגרפיות וחלופות נסיבתיות. משום כך עלינו לאסוף נתונים על פי שני סוגי קריטריונים:

- קריטריונים דמוגרפיים
- קריטריונים לבחינת נסיבות השיח

8. לדוגמה, המאגר הלאומי הבריטי (The British National Corpus), שבו מאה מיליון מילה, כולל תת-מאגר של עשרה מיליון מילה של לשון מדוברת, והוא מחולק לשני חלקים זהים: חלק החלופות הדמוגרפיות, ובו תמלילי הקלטת שיחות טבעיות של כלל הציבור, וחלק החלופות הנסיבתיות, ובו תמלילי הקלטות שהוקלטו בנסיבות מגוונות. ראה <http://info.ox.ac.uk/bnc/what/balance.html>; השווה ברגלונד, 1999, § 2.1, עמ' 31-32; וכן § 5.1.2 להלן.

מעמ"ד יורכב משני תת-מאגרים: מאגר ראשי ומאגר משלים. המאגר הראשי יהווה את חלק הארי של מעמ"ד ויכלול כ-90% מכלל הנתונים. המאגר המשלים שבו שני חלקים, דמוגרפי ונסיבתי, יכלול כ-10% מהנתונים. הרכב מעמ"ד:

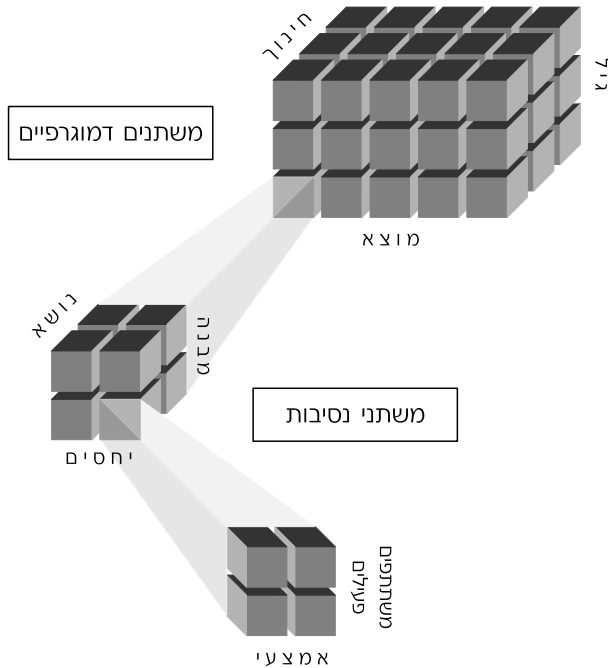
ממה	ממד	מר
-----	-----	----

מר = מאגר ראשי (90%); ממד = מאגר משלים דמוגרפי (5%); ממה = מאגר משלים נסיבתי (5%)

המאגר הראשי יורכב מתאים שייבנו על פי קריטריונים דמוגרפיים משולבים בקריטריונים לבחינת נסיבות השיח. נשתמש בתבנית שציריה העיקריים יהיו ציר דמוגרפי וציר נסיבות השיח.



כל אחד משני הצירים מהווה כשלעצמו מרחב רב-ממדי. במודל זה כל אחד מארבעים וחמישה התאים בתבנית הדמוגרפית (חמישה משתני מוצא מוכפלים בשלושה משתני גיל ומוכפלים בשלושה משתני השכלה) בנוי להכיל שמונה תאים שונים על פי משתנים נסיבתיים (שני משתנים של יחסים בין-אישיים מוכפלים בשני משתנים של מבנה שיח ומוכפלים בשני משתנים של נושאי שיח). כל תא מתאים אלה עשוי להכיל עוד ארבעה תאים (מונולוג או דיאלוג, שיחת פנים-אל-פנים או שיחה טלפונית). פרטים נוספים על אודות המבנה הזה יידונו בהמשך (להלן, §5.1).



הנתונים במאגר המשלים ייערכו בשני תת-מאגרים שונים. תת-מאגר אחד ייועד להשלמת המאגר הראשי ויושתת בעיקרו על קריטריונים דמוגרפיים, והוא ידגום בדגימה לא-מאוזנת קבוצות אוכלוסין אשר ייצוגן יהיה חסר במאגר הראשי. תת-מאגר האחר יושתת בעיקרו על קריטריונים של נסיבות שיח ויכיל דגימות טקסטים מאמצעי התקשורת, מן הכנסת ומבתי המשפט. כל תת-מאגר מן המאגר המשלים יורכב מחמישים תאים, דהיינו 5% מכלל מעמ"ד. מבנה המאגר המשלים יפורט ב-6 להלן.

מעמ"ד על שני חלקיו, הראשי והמשלים, יאוחסן בבסיס נתונים רחב ומתוחכם. כל חוקר יוכל לשלוף את הנתונים על פי משתנים כחפצו, וגם לפי שילובים מגוונים של משתנים. לדוגמה, אם ירצה חוקר להתמקד במאפייני לשונם של דוברים ילידים ממוצא צפון אפריקני, צעירים ובעלי השכלה על-תיכונית, הוא יוכל לקבל את המידע על פי הקריטריונים האלה.

4.2 דגימה וכלי ניתוח

מחובתנו להדגיש כאן את ההבחנה בין הקריטריונים שישמשו לדגימה לבין הקריטריונים שישמשו לניתוח החומר הלשוני. האתגר העיקרי בתיעוד מדגמי של

אוכלוסיה המונה ארבעה מיליון נפש ויותר⁹ לצורך מחקר שימושי הלשון וחלופותיה הוא להגיע לייצוג רחב ככל האפשר של חלופות הלשון, לתעד טקסטים באורך מספיק לחקירתם, ועם זאת לשמור על ממדים סבירים לעיבוד הנתונים במאגר. המשימה מורכבת עוד יותר בשל העובדה שהנתונים הדמוגרפיים כשלעצמם נערכים על פי רובד נוסף, הוא רובד נסיבות השיח על מאפייניו השונים. בעוד שהמאגר יכול בבסיסו נתונים טקסטואליים שייאספו על פי בדיקה מדגמית אקראית של שכבות האוכלוסיה, מידע דמוגרפי מפורט ומידע לגבי נסיבות השיח, שיילקט עם ביצוע ההקלטות ויאוחסן במאגר המידע, יאפשר עיבוד סוציו-לינגוויסטי ולשוני מורכב הרבה יותר.

דגימת האוכלוסין עצמה תיעשה על פי קריטריונים סטטיסטיים כדי לשקף כמותית את כלל האוכלוסיה (ראה §5.1.2). לעומת זאת, איסוף נתונים המתבצע על פי קריטריונים אנליטיים לאו דווקא יהא מייצג מבחינה סטטיסטית. במילים אחרות, המשתמש בבסיס הנתונים של המאגר יוכל לשלוף נתונים לשוניים מתוך בסיס נתונים, שאמנם יהיה ערוך על פי קריטריונים סוציו-לינגוויסטיים, אולם לא בכל מקרה יהוו הנתונים שיוצאו ייצוג כמותי של האוכלוסיה המיוצגת בטקסטים אלה. כפי שיוסבר בהמשך, מעמ"ד שואף לגשר בין אין-ספור חלופות השיח הנהוגות בקרב קהילת הדוברים של העברית הישראלית לבין ייצוגה במאגר על ידי אפיון השונות במונחים דמוגרפיים ונסיבתיים בהתאם לקריטריונים סטטיסטיים ואנליטיים. עריכת מדגמים היא תחומם של סטטיסטיקאים ושל סוציולוגים. המידע על תהליך עריכת המדגם, עם כל חשיבותו בעת תכנון המאגר, אינו רלוונטי למשתמש.¹⁰ לעומת זאת, הקריטריונים שיינקטו בעת העיבוד המדעי של הטקסטים, דהיינו על ידי החוקרים שיעשו שימוש בנתונים, חייבים להיות כפופים לקריטריונים הנהוגים במחקרים בלשניים וסוציו-לינגוויסטיים. כפי שיוסבר להלן, הנתונים למאגר הראשי במעמ"ד ייאספו בדגימה אקראית ולאחר מכן ייערכו לטובת המשתמשים בהם בהזנת התבנית הרב-תאית מתוך הקפדה על הקריטריונים האנליטיים. הקריטריונים האנליטיים לגבי הנתונים הסוציו-לינגוויסטיים במאגר יהיו שונים ורבים יותר מאלה שישמשו אותנו בעת כינון המאגר וארגונו. כדי להקל על המשתמש ולאפשר הסקת מסקנות הנשענות על בסיס איתן ובלא להפר את עקרון הייצוגיות יהא עלינו להגדיר כמה קריטריונים אנליטיים, הן דמוגרפיים והן בלשניים, עוד בשלב ארגון החומר.

9. האוכלוסיה הנדגמת אינה כוללת ילדים עד גיל 15 (ראה §5.1.1 להלן).
 10. לדגימת אוכלוסיה לניתוח בלשני ראה מילרוד, 1987, פרק 2. כמוכן, כל שינוי או גיוון שיתבקשו לאחר עריכת הדגימה האקראית יפורסמו ויובהרו. כדי למנוע ככל האפשר הטיות בתוצאות הדגימה, תיערך הרצה סטטיסטית מקדמית. לפרטים נוספים ראה אתר מעמ"ד < <http://www.tau.ac.il/humanities/semitic/maamad.html> > .

5. המאגר הראשי

5.1 קריטריונים אנליטיים

המאגר יכלול חמישה מיליון מילה – אלף תאים בני חמשת אלפים מילה כל אחד. המאגר יורכב אפוא מאלף תאים של טקסט מוקלט.

"תא" הוא היחידה הסוציולוגוויסטית הבסיסית במעמ"ד. התא הוא קטע דיבור מוקלט ובו חמשת אלפים מילה של טקסט רציף ומלוכד או כמה קטעים שכוללים יחדיו חמשת אלפים מילה, כגון טקסט יחיד רציף ובו חמשת אלפים מילה של יהודי אשכנזי בן 20 בעל השכלה תיכונית שהוקלט בעת הרצאת דברים מסודרת. תא עשוי להיות מורכב גם משני טקסטים רציפים בני אלפים ושלושת אלפים מילה בהתאם, ובהם שיחה בין שני אנשים. דוגמה אחרת: תא שבו שיחה רציפה בת חמשת אלפים מילה בין כמה אנשים בעלי רקע סוציולוגוויסטי שונה (דוגמה לטקסטים מתא כזה מובאת במאמרו של ש' יזרעאל בחוברת זו, עמ' 289–314).

המאגר הראשי, שיהווה 90% ממעמ"ד, יכיל תשע מאות תאים שיידגמו בדגימה אקראית. מלבדם ייאספו נתונים לעוד חמישים תאים שלא יידגמו אקראית לשם ייצוג קבוצות מיוחדות שאנחנו מניחים שהן משמעותיות למצב הלשוני בישראל. חמישים התאים הבלתי מאוזנים או הבלתי משוקללים שבתת-מאגר זה ייתוספו לתשע מאות התאים המאוזנים ויהוו חלק ממאה התאים של המאגר המשלים (ראה § 6.1 להלן).

5.1.1 קטגוריות דמוגרפיות

כבואנו לדגום אוכלוסיה עבור מחקר בלשוני או סוציולוגוויסטי אנו מלקטים נתונים סוציולוגוויסטיים על הדוברים. המידע חייב לייצג את השונות הקיימת בקהיליית הדוברים בכללותה. שונות זו נובעת מהבדלים במקום היוולדם של הדוברים, מהיותם דוברים ילידים או לא-ילידים של השפה, מהעדה שאליה הם שייכים, מסוגו של מקום מגוריהם (עירוני, כפרי, קיבוץ וכדומה), מהגיל, מהמין, מהמעמד החברתי-כלכלי, מהמקצוע, מהעיסוק, מהשירות הצבאי, מהדת, משהייה ממושכת שלהם בחו"ל ומהשפות המדוברות בבית. הדגימה תיערך על פי הקריטריונים הסטטיסטיים המתחייבים (§ 4.2) מתוך היצמדות למטרת כיוון מאגר המכיל חמישה מיליון מילה שיאפשר גישה לעיבוד נתוניו ושליטה במידע הכלול בו. עם כל אחד מהאינפורמנטים שיוקלטו ייערך ריאיון סוציולוגוויסטי על מנת להפיק את מרבית המידע הסוציולוגי הרלוונטי על אודותיו, מידע שישמש מחקרים בלשוניים וסוציולוגוויסטיים נוספים. כל הקריטריונים שפירטנו לעיל ייכללו בבסיס הנתונים הסוציולוגוויסטיים ויהיו נגישים לכל דורש. מחלוקת יתר של המדגם לתת-קבוצות רבות מדי יתקבל אוסף של אידיולקטים ולא דווקא מאגר מייצג ונגיש של כלל קהילת הדוברים. על כן נצמצם את מספר המשתנים שלפיהם יאורגנו נתוני המאגר על מנת לאפשר אחזור מידע מייצג בר-כימות של דוברי העברית.

הגישה שנקטנו לכינון המאגר המייצג היא גישה תלוית תרבות. תבנית המאגר הראשי תוכננה כך שתתאים למבנה הייחודי של החברה הדוברת עברית. בעת תכנון המאגר ראינו לנגד עינינו את תבנית החברה הישראלית כמכלול של מקטעים העשויים להיחשב כקהילות דוברים (אף שלעת עתה אין אנו יכולים אלא להעלות השערות באשר לדומה ולמפריד ביניהן). מובן מאליו כי השונות מעוגנת במאפיינים דמוגרפיים¹¹ ועלינו לנתחם כראוי. הנחת העבודה שלנו מתבססת על שלושה קריטריונים דמוגרפיים שנואים לנו החשובים ביותר ביצירת השונות הלשונית בישראל: (א) עדה או דת, מקום לידה וארץ מוצא המשפחה; (ב) גיל; (ג) השכלה.

א. עדה או דת, מקום לידה וארץ מוצא המשפחה (חמש קטגוריות):

1. יהודים, ילידי הארץ, בני אב יליד אסיה או אפריקה¹²

2. יהודים, ילידי הארץ, אחרים

3. יהודים, ילידי חו"ל, שעלו לפני 1965

4. יהודים, ילידי חו"ל, שעלו אחרי 1965

5. לא-יהודים (מוסלמים, נוצרים, דרוזים)

חלוקה זו גסה למדי, אבל היא מאפיינת את הרכב החברה בקטגוריות הגדולות ביותר; ראשית, בהבחנה בין יהודים לבין לא-יהודים (קטגוריות 1-4 לעומת קטגוריה 5); שנית, בהבחנה בין מה שנחשב ליסוד המבדיל הגדול בין יהודים בחברה – בין אשכנזים לבין לא-אשכנזים (משתנים 1-2).¹³ עוד מבחינה חלוקה זו בין דוברים ילידים ולא-ילידים של העברית (משתנים 1-2 לעומת 3-4), ואת הדוברים הלא-ילידים החלוקה ממיינת בהתאם למועד עלייתם לארץ (משתנים 3-4), כלומר דוברים שהשנים המכריעות בהשכלתם עברו עליהם או לפני הקמת המדינה ובימי ביסוסה לאחר 1948 או משנת 1965 ואילך – על פי גלי העלייה השונים.¹⁴

11. קהילות דוברים מתאפיינות גם בשוני בנסיבות השיח הנהוגות בהן (ראה בייבר, 1995). מעמ"ד יבחין בין נסיבות שיח על ידי מאפיינים מושגיים ולא דווקא על ידי פירוט הנסיבות האפשריות עצמן (ראה 5.1.2§ להלן).

12. הסיבה להסתמכות על מקום לידת האב היא טכנית: כך נוהגים מפקדי האוכלוסין כדי לאפשר השוואת נתונים, אימצנו זאת אף אנו.

13. המשתנים הנקוטים כאן משמשים אך ורק את הבדיקה הסטטיסטית, כפי שהסברנו בערה 4 לעיל. ההבחנה בין אשכנזים לבין לא-אשכנזים (או ספרדים) שימשה בחקר הלשון בשל זיהוי העברית הישראלית האשכנזית כעברית כללית לעומת העברית המשוערכת שבפי הדוברים ממוצא ספרדי ומזרחי. כך אצל בלנק, 1956א, עמ' 189; בלנק, 1956ב; ברמן, 1997, עמ' 312-313; בולוצקי, 1997, עמ' 287). אנו מניחים שדואליות זו כוללנית מדי בבואה להעיד על חלופות אמתיות בישראל היום (השווה דבנס, 1980; 1981).

14. ב-1948 מנתה האוכלוסייה היהודית רק 650,000 איש. 684,000 העולים שהגיעו לארץ בין השנים 1948 ו-1951 הכפילו את האוכלוסייה, ובשנות השישים המוקדמות הגיע מספר היהודים בארץ לשני מיליון.

ב. גיל (שלוש קטגוריות):

1. צעירים (בני 15–27)

2. בוגרים (בני 28–50)

3. מבוגרים (בני 51 ומעלה)

מעמ"ד ידגום את האוכלוסיה בישראל החל מגיל 15, שהוא גיל המעבר לחינוך על-יסודי. האוכלוסיה הנדגמת תחלק לשלוש קבוצות גיל בהיקף דומה כדי להבטיח תוצאות משמעותיות מבחינה כמותית. בעת ניתוח הנתונים יהיה אפשר לחלק את קבוצת הגיל הראשונה לשתי תת-קבוצות, מתוך הנחה שקיים שינוי לשוני משמעותי בגיל סיום בית הספר התיכון והגיוס לצבא או בתחילת לימודי המשך (באשר להשפעת השירות הצבאי על התרבות הלשונית בישראל ראה בהמשך, §6.1). משום כך נציע להפריד הפרדה נוספת בין מתבגרים עד גיל הגיוס (גילאי 15–18) לבין צעירים (גילאי 19–27). תחילת קבוצת הגיל (2) נקבעה בהתאם לגיל הממוצע של נישואין, 27, גיל שבו מתחילים הצעירים להקים משפחה, וסיומה נקבע בגיל שבו ילדיהם גדולים ומתחילים לעזוב את הבית. בדומה להצעתנו לחלוקה נוספת של קבוצת הגיל הראשונה נציע לחלק גם את הנדגמים בני 51 ומעלה (קבוצה 3) לשתי תת-קבוצות: עד גיל הפרישה ולאחריו.

ג. השכלה (שלוש קטגוריות)

1. מי שלא סיים בית ספר תיכון

2. בעלי השכלה תיכונית

3. בעלי השכלה על-תיכונית

אם כן, הציר הדמוגרפי של תבנית התאים הרב-ממדית שתשמש כבסיס לניתוח מעמ"ד הוא עצמו רב-ממדי וכולל שילוב הקטגוריות של משתני העדה או הדת, מקום הלידה וארץ מוצא המשפחה עם משתני הגיל ועם משתני ההשכלה. קריטריון חשוב נוסף הוא מין הדוברים. תבנית התאים המוצעת אינה כוללת נתון זה כקריטריון נפרד משום שמין הדוברים יובחן בבירור בעת הדגימה, ואנו צופים כי החלוקה בין גברים לבין נשים בכל קטגוריה אנליטית תהיה בהתאם לחלוקת המינים באוכלוסיה, דהיינו 1:1.

5.1.2 קטגוריות נסיבות השיח

היות שהלשון הננקטת בעת הדיבור תלויה בנסיבות השיח, על מאגר של שפה מדוברת לשקף סוגים שונים של נסיבות. ארגון תבנית נסיבות השיח עבור מעמ"ד מושתת על העיקרון התולה את נסיבות השיח בשלושה גורמים מכריעים: היחסים בין בני השיח, ארגון השיח ונושא השיח. שלושת הגורמים האלה (משתנים א–ג) הם העיקריים בחמשת המשתנים שעל פיהם בנויה תבנית נסיבות השיח. השניים הנוספים הם טכניים יותר: מספר המשתתפים בשיח וצינור התקשורת (משתנים ד–ה).

משתנים עיקריים

א. יחסים בין-אישיים: קרבה מול ריחוק (+/-קרבה) במשתנה (א) משתקפים היחסים האישיים. יחסים בין בני משפחה או בין חברים קרובים יוגדרו כיחסי קרבה (+קרבה).

ב. מבנה השיח: תלוי-תפקיד מול שוויון (+/-תפקיד) במשתנה (ב) משתקף מבנה השיח. מבנה שיח שבו חלוקת תפקידים ברורה בין בני השיח (למשל, כשיש תפקיד סמכותי למי מהם) יובחן מול שיח שבו אין חלוקת תפקידים כזו (+תפקיד).

ג. נושא השיח: אישי מול לא-אישי (+/-אישי) במשתנה (ג) משתקף תוכן השיח, ובו הבחנה בין נושא שיחה אישי לבין נושא שיחה לא-אישי.

שלוש הקטגוריות שנמנו (א-ג) יבואו לידי ביטוי במאגר בכל שמונת הצירופים האפשריים (2³ שילובים) כמפורט להלן. לעומת זאת, המשתנים המשניים דלהלן (ד-ה) ייושמו רק בחלק מהתבנית, בהיותם שכיחים הרבה פחות בקרב רוב קהילות הדוברים.

משתנים משניים

ד. משתתפים פעילים: מונולוג מול דיאלוג (+/-מונולוג) על אף שמונולוג כשיח של אדם יחיד עשוי להימצא בכל אחת מההקלטות, נבחין במשתנה זה רק במקרים המובהקים שבהם מתפתח מונולוג, כלומר כשהמונולוג מגדיר מהותית את סוג השיח.

ה. ערוץ: טלפון מול פנים-אל-פנים (+/-טלפון) גם שיחה באמצעות הטלפון תובחן כמהותית רק במקרים שיאופיינו ככאלה.

ארגון המשתנים הנסיבתיים

טלפון	מונולוג	אישי	תפקיד	קרבה	
-	-	+	-	+	1
+	-	+	-	+	ב1
-	-	-	-	+	2
+	-	-	-	+	ב2
-	-	+	+	+	3
-	-	-	+	+	4
-	-	+	+	-	5
-	+	+	+	-	א5
-	-	-	+	-	6
-	+	-	+	-	א6
+	-	-	+	-	ב6

קרבה	תפקיד	אישי	מונולוג	טלפון
7	-	+	-	-
8	-	-	-	-

הערה: א לצד מספר משמעו מונולוג; ב לצד מספר משמעו שיחה טלפונית

דוגמאות לסימול קטגוריות על פי המשתנים:

- 1 משפחה או חברים – שיחת יום-יום
- ב1 משפחה או חברים – שיחת יום-יום טלפונית
- 2 משפחה או חברים – שיחה שלא בנושא אישי (על פוליטיקה, למשל)
- ב2 משפחה או חברים – שיחה טלפונית שלא בנושא אישי
- 3 משפחה בעלת גינונים מסורתיים – שיחת יום-יום
- 4 משפחה בעלת גינונים מסורתיים – שיחה שלא בנושא אישי (על פוליטיקה, למשל); שיעור בלתי פורמלי באוניברסיטה
- 5 פגישה טיפולית; התייעצות עם רב
- א5 פגישה טיפולית; הקראת סיפור
- 6 פגישת עסקים; ראיון עבודה
- א6 הרצאה רבת משתתפים באוניברסיטה; נאום פוליטי
- ב6 ראיון עבודה טלפוני; שיחת עסקים טלפונית
- 7 בחדר ההמתנה במרפאה
- 8 שיחה אקראית שלא בנושאים אישיים בין שני קונים במרכול

כאמור, תבנית המדגם למאגר הראשי במעמ"ד היא רב-ממדית. מאגר אידאלי, אוטופי, היה כולל ייצוג דמוגרפי של כל האוכלוסיה ובכל נסיבות השיחה האפשריות. מאחר שאידאל כזה אינו בנמצא, עיברנו מנגנון שבו החלופות הנסיבתיות שקולות כנגד מספר הדוברים והשפעתם הפוטנציאלית על הלשון ועל תופעות לשוניות.

כפי שהוסבר לעיל, הציר הדמוגרפי כשלעצמו הוא רב-ממדי וכולל ארבעים וחמישה צירופים (חמישה צירופים של משתני עדה או דת, מקום לידה וארץ מוצא המשפחה מוכפלים בשלושה משתני גיל ומוכפלים בשלושה משתני השכלה). כל אחד מצירופים אלה הוא בעל פוטנציאל להכיל שמונה חלופות נסיבתיות, כשכל חלופה בעלת פוטנציאל להכיל ארבע חלופות נוספות (מונולוג או דיאלוג, שיחת פנים-אל-פנים או שיחה טלפונית). היות שחלק מהחלופות הנסיבתיות אינן פוריות בצירופים דמוגרפיים מסוימים,¹⁵ נכניס שינויים לתבנית כולה על ידי מתן משקל יתר לחלופות המשותפות לכמות גדולה יותר של דוברים ולחלופות שנראות כמשפיעות יותר על חייה הלשוניים של קהילת הדוברים.

15. לדוגמה, קרוב לוודאי ששיח של יהודי מבוגר יליד הארץ שאביו נולד באסיה או באפריקה והוא בעל השכלה בסיסית בלבד לא ייקלט בחלופה הנסיבתית מס' 1, משום שסביר להניח כי ישתייך למשפחה מסורתית.

התרשים הבא מייצג את השפעתן היחסית של החלופות השונות ומספר היקריותיהן במאגר. בעוד שהחלופות בעלות ההשפעה הגבוהה ביותר ייוצגו במעמ"ד על ידי ארבעה תאים לחלופה, לחלופות בעלות ההשפעה הנמוכה ביותר לא יוקצו תאים בתכנון המאגר הראשי.

חשיבות	חלופות	מספר תאים לחלופה	סך כל התאים במאגר
מרבית	3,1	4	8
רבה	6,6,4,2	2	8
מועטה	6,א5,ב2,1	1	4
מזערית	8,7,5	–	–

מכאן שכל אחת מארבעים וחמש החלופות הדמוגרפיות במאגר תכיל עשרים חלופות נסיבתיות. כלל צירופים אלה הם תשע מאות התאים המהווים את המאגר הראשי במעמ"ד.

אנו מניחים כי הן מן הבחינה התאורטית והן מן הבחינה המעשית אלה הם המשתנים המבחינים בין סוגי השיח העיקריים. יחידות שיח שיובחנו על פי משתנים אלה ישקפו מבנה לשוני שונה. את ההנחה הזאת נבדוק במחקר המקדים ובמהלך איסוף הנתונים.

5.2 דגימה

מטרת כינון תאים מייצגים מבחינה סטטיסטית היא איסוף נתונים מספיקים למחקר סוציו-לינגוויסטי או בלשני. אנו נשתדל לכונן מאגר שגם יהווה תמונת המערכת הלשונית בכללה וגם ייצג נאמנה מגוון רחב של שונות לשונית. לשם כך אנו מציעים להשתמש במדגם מייצג של השפה המדוברת וחלופותיה. עקרונית, ניתן היה להשיג זאת על ידי דגימה של כל אחד מהתאים בתבנית שלנו והמכלול היה משקף את השפה כפי שהיא מדוברת, אולם למעשה אין הדבר אפשרי. דוברים וקטעי שיח אינם באים, בדרך כלל, מתויגים על פי תאי שיוך. יתר על כן, איננו יודעים לאיזה יסוד יש משקל יתר בתיוג – למאפייני הדובר או לנסיבות השיח. האפשרות שנבחרה היא דגימה בשני שלבים: קודם כול דגימת דוברי העברית בארץ, ולאחריה דגימת מאפייני השיח של כל דובר.

5.2.1 דגימה דמוגרפית

ביצירת מדגם אוכלוסין לצורך זה או אחר אפשר לחלק את האוכלוסיה לתת-קבוצות ולדגום כל אחת מהן בנפרד ("ריבוד" הדגימה). ריבוד דגימה עשוי להידרש מכמה טעמים, והחשובים בהם הם:

1. לתת-קבוצות מסוימות מאפיינים השונים במידה רבה מאלה של כלל האוכלוסיה (בעיקר אם הקבוצה קטנה יחסית או אינה מעורה דייה באוכלוסיה);

2. דווקא על שום ממדיה המצומצמים של קבוצה מסוימת, חשוב שהמדגם ישקף אותה ביחס נכון לממדיה;

3. הקבוצה קטנה וברצוננו לתת לה ייצוג מוגבר.

כינון מדגם ללא ריבוד אין פירושו בהכרח שתוצר הדגימה לא יהיה מייצג דיו או שיש בו ערך פחות מאשר במדגם מרובד. היעדר הריבוד משמעו שאנו מניחים לתהליכים הסטטיסטיים האקראיים לייצג את האוכלוסיה הנדגמת במידה סבירה. לפיכך דגימה אקראית של המאגר לפי אזורים מגורים תשמש מדגם ראשוני בבחירת האינפורמנטים.

חלקים נרחבים באוכלוסיה אכן יתועדו בכמות ניכרת, ואנו צופים שיעטבר די חומר לייצוג לשוני הולם. עם זאת, רמת ייצוגן של קבוצות אוכלוסין מסוימות לא תספיק לשם הסקת מסקנות כלליות על מאפייני לשונה של קבוצה זו. דוגמה לקבוצת אוכלוסין כזו היא אוכלוסיית תושבי הקיבוצים, המהווה שני אחוזים מכלל האוכלוסיה. חלוקה נוספת של בני הקיבוצים על פי גיל ומין, לדוגמה, תניב אך בודדים בכל יחידת דגימה. למשל, בנות קיבוצים בגילאים הצעירים של לפי השירות הצבאי מהוות אך 1.1% מאוכלוסיית המדינה שתידגם לצורך כיון מעמ"ד (קצת למעלה מארבעה מיליון נפש, כמוסבר לעיל). היות שבמאגר 900 תאים, הרי אפילו מדגם אידאלי יכול רק בת קיבוץ אחת מקבוצת גיל זו. דגימה כזאת היא כמובן מצומצמת מכדי שתוכל לאפיין כלשנית את שפת בנות הקיבוץ הצעירות, בייחוד כשאנו שואפים לכלול יותר מחלופה נסיבתית יחידה של כל תת-קבוצה. אף על פי שמעמ"ד יכול לייצג מסוים ללשוניותה של תת-קבוצה כזאת, התא או התאים שייכללו בו יוכלו רק לכוון את החוקרים באשר לטיפול הדרוש למחקר מקיף מן הסוג הזה אך לא לייצג נאמנה. במחקרים מן הסוג הזה יש לאסוף נתונים מאוכלוסיית היעד בנפרד.¹⁶ החוקרים יוכלו להשתמש במעמ"ד כבמקור להשוואה, כמסגרת התייחסות לגבי כלל האוכלוסיה, או, מה שחשוב יותר, כדי לקבל מידע ראשוני באשר לסוג המחקר הנחוץ לכל אחת מאוכלוסיות היעד או קהילות הדוברים המסוימות האלה. באשר לקבוצות הלשוניות הגדולות, המאגר בן חמישה מיליון המילה יספיק לעריכת מחקרים מקיפים על מגוון רחב של היבטים בלשניים וסוציו-לינגוויסטיים של העברית המדוברת בישראל.

5.2.2 דגימת נסיבות השיח

עריכת מדגם מייצג על פי נתונים דמוגרפיים היא, כאמור, טכניקה שכיחה הנהוגה בדגימת אוכלוסין ותחבצע על ידי דגימה אקראית של אוכלוסיית ישראל. לעומת זאת, הניסיון שנצבר בבניית מאגר לשון שייצג נאמנה גם את נסיבות השיח מועט ביותר.¹⁷

16. במקרים אחדים ייתכן שאפשר יהיה לתקן במידת-מה את בעיית תת-הייצוג באמצעות חמישים התאים של החלק הדמוגרפי של המאגר המשלים (ראה § 6.1 להלן).

17. לשאלת ייצוג נסיבות השיח ראה למשל קראודי, 1993, עמ' 262-263; ברגלונד, 1999, § 2.1;

כדי להבטיח ייצוג אמין של השונות הלשונית, הדמוגרפית והנסיבתית כאחת, יישמר עקרון האקראיות בעת דגימת כל הנתונים הטקסטואליים. דגימת נסיבות השיח תבצע לאחר איסוף כל ההקלטות של האוכלוסיה הנדגמת דגימה דמוגרפית. כל מי שעלה במדגם האקראי יתבקש להקליט את כל מהלכיו ואת הקורות אותו במשך זמן מסוים, כגון, במשך יום או יומיים רצופים. פרק הזמן הקצוב יתחלק שווה בשווה בין כל האינפורמנטים. אידאלית, יוקלטו שבעה רצפים שווים כאלה, וכל אחד מהם יפתח ביום אחר בשבוע. כל הקלטה כזאת תיבדק ויוסרו מרווחי שתיקה וקטעי דיבור שאינם ברורים. מכל החומר הנקי ייבחר באקראי קטע של טקסט הנמשך שעה, והוא שיהווה מקור לנתונים שייכללו במאגר הראשי. בחירה אקראית זו עשויה לשקף את ביזור מצבי השיח הטבעיים, הן על פי היקריותיהם בזמנים שונים, הן בהתאם לנסיבות השיח והן בהתאם לסוגי השיח המזומנים לסוגי אוכלוסיה שונים. דגימה זו, בשל אקראיותה והיקפיה, אינה תואמת בהכרח את המודל התאורטי שעל פיו נבנתה תבנית התאים. אכן, תבנית התאים הרב-ממדית נבנתה לצורך מענה אנליטי לחוקרי הלשון והחברה. אם כן, היא עלינו לתת את הדעת לעניין גישור הפער בין תביעות הסטטיסטיקה לבין הקריטריונים האנליטיים.¹⁸

5.3 גישור בין תביעות הסטטיסטיקה לבין הקריטריונים האנליטיים: היערכות התבנית הרב-מדית

התכנית הבסיסית של מעמ"ד מקצה עשרים תאים לכל חלופה דמוגרפית. היות שיש ארבעים וחמש חלופות דמוגרפיות, עולים משילובן תשע מאות תאים. כצפוי, לא כל החלופות תואמות זו את זו. למשל, קריטריון מקום הלידה 3 (יהודי שאינו יליד הארץ ושעלה לפני 1965) אינו תואם את קבוצת הגיל הראשונה (15–27 שנים) ואף לא את כל קבוצת הגיל השנייה (28–50 שנים), שגם בה יש ילידי התקופה שאחרי 1965. בנוסף על כך, חלק מהנדגמים בקבוצת הגילאים הצעירים אינם יכולים

> http://info/ox/ac/uk/bnc/what/spok_design.html ; ביבר, 1995, § 3 ; קנדי, 1998,

עמ' 71 ; אטקינס, קליר ואוסטלר, 1992 ; מקרתי, 1998, § 1.4. סיכום קצר על שאלת הייצוג

התאורטי ראה אצל יזרעאל, הרי ורהב, 2002.

18. עלינו להעלות בהקשר זה את שאלת תפוקת הדיבור השונה של דוברים שונים, שהרי יש המרבים לדבר ויש הממעטים. בהתחשב בכך, עלינו להכריע בשאלה אם להתמקד בייצוגיות הדוברים או בייצוגיות השיח. אם ננהג על פי עקרון כמות המילים השווה לכל דובר ללא זיקה יחסית לדיבור שהופק ממנו, יתקבל ייצוג הולם לדוברים, אך הטקסטים של אלה שהפיקו רצפים ארוכים יותר ילקו בתתי-ייצוג. לעומת זאת, אם ניצמד לייצוגיות השיח, הטקסטים של הממעטים בדיבור לא ייוצגו ביחס הולם. משום כך – ותהא האסטרטגיה שתינקט אשר תהא – ייערך רישום פרטני מלא של כל דובר על מנת לעמוד על דגמי השיח של הדוברים או של הטקסטים המופקים על ידי שקלול פשוט. היות שאנו שואפים לכוון מאגר מייצג של הלשון ושל השונות הלשונית יותר מאשר לתאר הרגלי שיח שכוחים, אנו מתכננים להיצמד לאפשרות הראשונה תוך כדי משקל שווה לדגמי השיח המשמשים בפי הדוברים.

להיות בעלי השכלה על-תיכונית (משתנה ההשכלה מספר 3). משום כך, כבר על פי התכנון הראשוני לא תוכל התבנית להתמלא בכל תשע מאות תאיה. כמו כן אנו סבורים שלא כל החלופות האפשריות מבין החלופות הנסיבתיות זמינות וקיימות עבור כל אחת מן החלופות הדמוגרפיות.

כפי שהבהרנו לעיל, איסוף הנתונים להצבתם במאגר וליקוטם בעת השימוש המחקרי בהם יתבצעו בדרכים שונות לחלוטין. ראשית יתנהל איסוף הנתונים להצבתם במאגר. משימה זו תבוצע באמצעים סטטיסטיים מוקפדים, כלומר על פי דגימה אקראית.

ארגון החומר הטקסטואלי שנאסף ייעשה בשלב הבא. עריכת הנתונים לשימוש מחקרי תבסס על העקרונות שנקבעו בכינון תבנית התאים על פי הקריטריונים הדמוגרפיים והקריטריונים הנסיבתיים ששוקללו כפי שתואר לעיל. יעמדו לרשותנו טקסטים שנבחרו באקראי מתוך מצאי טקסטים שהופקו על ידי אנשים שנבחרו באקראי ובמגוון רחב של נסיבות שיח. הטקסטים הללו ייערכו בתבנית הרב-ממדית, וכל יחידת טקסט תוכנס אל התא התואם לה על פי הנתונים הדמוגרפיים של דובריה ועל פי נסיבות השיח שלה. חלוקת החומר הטקסטואלי לתאים תיעשה על פי עקרון האקראיות ובכימות הולם. דהיינו, תבנית התאים תתמלא במצאי הטקסטים שהעמידה לרשותנו הבחירה האקראית ולא על פי מספר התאים שהעמיד העיצוב העקרוני לרשות כל יחידה דמוגרפית. גישה זו תאפשר לנו לעמוד על מגוון דגמי השיח הקיים בקהילת הדוברים וללמוד על אודות הדגמים מבחינה כמותית ומבחינה איכותית, כפי שיעלה משילוב הקריטריונים הדמוגרפיים והקריטריונים הנסיבתיים. במילים אחרות, המיון לתאים יניב ניתוח ראשוני של נסיבות השיח שאכן מצוי בשימוש בחלקי אוכלוסיה שונים וכן ניתוח ראשוני של היקף השימוש של החלופות הלשוניות הללו ביחס לחלופות אחרות. לדוגמה, תא עשוי להכיל טקסט יחיד בן חמשת אלפים מילה של הרצאה באוניברסיטה מפי דוברת ילידית בת 50 ממוצא מערב אירופי; או להכיל שני דיאלוגים פנים-אל-פנים בין חיילים בני 20 שנולדו ברוסיה, האחד בן אלפיים מילה והאחר בן שלושת אלפים מילה; או להכיל כמה שיחות טלפון קצרות בין מנהל לבין פקידיו. בכל המקרים כל אחד מהטקסטים השונים יהיה רצף אחד מלוכד. כפי שצוין לעיל, כל הנתונים הסוציו-לינגוויסטיים יהיו נגישים לכל דורש, וחוקר המאגר יוכל לשלוף טקסטים לצרכיו ועל פי נתונים סוציו-לינגוויסטיים כחפצו. עם זאת, כדי להניב תוצאות משמעותיות בניתוח לשוני, אנו מציעים את המשתנים שפורטו לעיל כבסיס לשליפת נתונים.

6. המאגר המשלים

המאגר המשלים יורכב משני תת-מאגרים שגודלם שווה, האחד על בסיס דמוגרפי, האחר על בסיס נסיבות שיח.

6.1 השלמת המאגר הדמוגרפי

צירופים מסוימים, דמוגרפיים או דמוגרפיים ונסיבתיים, לא יהיו מיוצגים בדגימה; יהיו גם צירופים שייצוגם יהיה זעיר מכדי לאפשר מחקר לשוני משמעותי. ברוב המקרים חסרים אלה אכן יהיו פועל יוצא של הריבוד הדמוגרפי של קהילת הדוברים בישראל או של מצאי נסיבות השיח הטבעיות לקבוצות השונות. בכל זאת, במקרים מסוימים יידרשו תיקונים, אם מפאת פגם כלשהו במדגם ואם בשל שיקול למתן ייצוג יתר לקבוצה כלשהי שנראה שהשפעתה מיוחדת על ההתנהגות הלשונית של דוברי העברית בישראל. עבור קבוצות אלה ייכוונו תאים שיוזנו בטקסטים שייבחרו בדגימה ל-אקראית. כפי שניתן לנבא בשלב זה, התאים האלה יכילו נתונים מדיכורם של בני קיבוצים, של חרדים או של דוברים ילידים של עברית ששהו תקופות ארוכות מחוץ לישראל.

לשון הצבא תזכה לתשומת לב מיוחדת. השירות הצבאי בישראל הוא חובה; גברים משרתים בשירות סדיר שלוש שנים ונשים – שנה ותשעה חודשים. לאחר מכן גברים ממשיכים ומשרתים בשירות מילואים במשך תקופה מסוימת, לעתים עד גיל 49.¹⁹ רבים אחרים משרתים בצבא הקבע או בכוחות הביטחון האחרים. בהיות מדינת ישראל מדינה קולטת עלייה במהותה שימש השירות הצבאי מאז ומעולם ככור ההיתוך של החברה הישראלית. על שום חשיבותו הסגולית של הצבא בחיי החברה בישראל השפעתו על העברית היא עצומה. השפעה זו ניכרת בעיקר באוצר המילים ובמטבעות לשון, אך אין ספק שהיא תורגש גם מעבר לתחומים אלה. ודאי שהמאגר הראשי יכיל בתוכו גם תת-מאגר של הקלטות מהצבא. בשלב זה אין לדעת אם תת-מאגר זה ייווצר במדגם האקראי או ייערך בנפרד.²⁰

6.2 מאגר משלים לנסיבות שיח

המאגר הראשי צפוי לתת ייצוג הולם לרוב סוגי הדיבור העיקריים. עם זאת, הן המאגר הראשי והן תת-המאגר הדמוגרפי במאגר המשלים אינם מייצגים תחומים אחדים של לשון הדיבור שחשיבותם רבה כדי כך, שמן הראוי שיימצא מקומם במעמ"ד. כך הוא דיבורם של חברי הכנסת בבית הנבחרים, הדיבור בכתי משפט, ומעל לכול לשון אמצעי התקשורת (טלוויזיה ורדיו). בהיות כלל האוכלוסיה חשוף למקורות אלה במידה זו או אחרת, חלופות לשוניות אלה, הגם שאינן משקפות את דיבורם של דוברים רבים, הרי הן בעלות השפעה רבה על הלשון. לשם כך ראינו לנכון להשלים את המאגר הראשי, שבעיקרו הוא מאגר מייצג, בתאים נוספים שישקפו את הלשון המדוברת בסביבות בנות השפעה לשונית כגון אלה. הקטגוריות

19. מעט נשים משרתות אף הן במילואים לפרק זמן קצר בהרבה מגברים.

20. הדבר תלוי באופן הדגימה שיינקט. דגימה מתוך מקומות מגורים תגרום לחסר משמעותי באינפורמנטים מתוך שורות הצבא, השווים זמן רב מחוץ לבתיהם.

המיוצגות בתת-מאגר זה מפורטות להלן. ייערך מעקב מתמיד אחר תוצר איסוף התאים הזה כדי לבדוק את הצורך להשלימו על ידי תוספת תאים מקבילים מפי דוברים בעלי רקע דמוגרפי שונה במובהק. הטבלה דלהלן מציגה את החלופות הנסיבתיות שיוצגו בתת-המאגר הנסיבתי.

דיבור לא-ספונטני		דיבור ספונטני			
טקסט מוקרא	חופשי				
		+	א1	שידור רגיל	טלוויזיה
	+		ב1	שידור רגיל	
+			ג1	שידור רגיל	
		+	א2	שידור ספורט	
	+		ב2	שידור ספורט	
+			ג2	שידור ספורט	
	+	+	3	ריאיון	
	+	+	4	רב-שיח	
+			5	סרט	
+			6	פרסומת	
		+	א7	שידור רגיל	רדיו
	+		ב7	שידור רגיל	
+			ג7	שידור רגיל	
		+	א8	שידור ספורט	
	+		ב8	שידור ספורט	
+			ג8	שידור ספורט	
	+	+	9	ריאיון	
	+	+	10	רב-שיח	
	+	+	11	שיחות טלפון	
+			12	פרסומת	
+			א13	נאום	כנסת
	+		ב13	נאום	
		+	ג13	מונולוגים; דיאלוגים	
+			א14	נאום	בית משפט
	+		ב14	נאום	
		+	ג14	מונולוגים; דיאלוגים	

תת-המאגר המשלים הזה, הבנוי בבסיסו על קטגוריות של נסיבות שיח, כולל עשרים ושישה תאים של חמשת אלפים מילה כל אחד, שהם מאה ושלושים אלף

מילה, כ-2.6% מן המאגר כולו. בנוסף, מספר התאים הללו יגדל בשל הצורך הוודאי להוסיף לתאים אלה תאים נוספים על פי נתונים דמוגרפיים שונים של הדוברים. עוד ייתוספו למאגר המשלים טקסטים שייצגו סוגות של מופעים מהתחום הלירי ומהתחום הדרמטי (המקורי והמתורגם) ושל מופעי יחיד וביצועים שונים, שהם בעלי השפעה לשונית בלתי מבוטלת על שימושי הלשון בישראל. הערכתנו היא כי תת-המאגר הנסיבתי יוכפל בגודלו, ובסופו של דבר יכלול כרבע מיליון מילה שיערכו בחמישים תאים ויהוו כ-5% מן המאגר כולו.

7. סיכום

מאמר זה מתאר את ארגונו הבסיסי של מאגר העברית המדוברת בישראל (מעמ"ד). למיטב ידיעתנו, זהו ניסיון ראשון לכינון מאגר מייצג על שני צירי משתנים, דמוגרפיים ונסיבתיים, בהתאם לקריטריונים סטטיסטיים ואנליטיים.

שילוב הקטגוריות הדמוגרפיות והנסיבטיות ישקף את תפוצת סוגי השיח בקבוצות שונות באוכלוסייה. הגישה הנקוטה כאן היא תלוית-תרבות מתוך כוונה ליישמה על המבנה הייחודי של החברה הדוברת עברית. ייוצגו דוברים ילידים ולא-ילידים של השפה; יבוא לידי ביטוי אופייה המיוחד של החברה על שום היותה חברה קולטת עלייה ובה מיעוט ערבי גדול, ותינתן הדעת על קהילות דוברים בעלות מאפיינים בלעדיים, כמו זו של הצבא.

בכוונתנו לאסוף את הנתונים למעמ"ד בתוך פרק זמן קצר כדי שתתקבל תמונת מצב סינכרונית של העברית הישראלית. אולם דגם העבודה המוצע יוכל לשמש לכינון מאגרים נוספים. בשינויים קלים יהיה אפשר להתאים את הדגם לאיסוף נתונים ממוקדים על קהילת דוברים מצומצמת בתוך החברה. תבנית התאים המוצעת בזה תוכל גם להוות בסיס לכינון מאגרים של העברית בעתיד. כך יעמוד מאגר העברית המדוברת בישראל לשימוש לא רק לשם מחקר העברית בתקופתנו, אלא גם למחקר תולדות התפתחות השפה, החברה והתרבות העבריות בישראל לאורך דורות. יש מקום גם להניח כי העקרונות שעליהם מושתת עיצוב זה של מאגר לשון – בהתאמות ושינויים המתחייבים – יוכלו לשמש אף לכינון מאגרי לשון של קהילות דוברים בשפות אחרות בחברות ובתרבויות מגוונות בעולם כולו.

הקיצורים הביבליוגרפיים

J. Edwards, "Survey of Electronic Corpora and Related = 1993 Resources for Language Researchers", *Talking Data: Transcription and Coding in Discourse Research*, ed. J. A. Edwards and M. D. Lampert, Hillsdale, New Jersey 1993, pp. 263–306

- S. Atkins, J. Clear and N. Ostler, "Corpus = 1992 קליר ואוסטלר, Design Criteria", *Literary and Linguistic Computing* 7 (1992), pp. 1–16
- F. Boas, *Race, Language and Culture*, New York 1940 = 1940 בואז,
- Sh. Bolozky, "Israeli Hebrew Phonology", *Phonologies of = 1997 בולוצקי, Asia and Africa (Including the Caucasus)*, 1, ed. A. S. Kaye, Winona Lake, Indiana, pp. 287–311
- D. Biber, "Methodological Issues Regarding Corpus-based = 1990 בייבר, Analyses of Linguistic Variation", *Literary and Linguistic Computing* 5 (1990), pp. 257–269
- D. Biber, *Dimensions of Linguistic Variation: A Cross- = 1995 בייבר, Linguistic Comparison*, Cambridge 1995
- בלאו, תשנ"א = ' בלאו, דקדוק העברית החדשה, לשוננו נה (תשנ"א), עמ' 157–149 [=ביקורת על גלינרט, 1989]
- בלאו, תשנ"ו = ' בלאו, "פתח דבר", הלשון העברית בהתפתחותה ובהתחדשותה: הרצאות לרגל מלאות מאה שנה לייסוד ועד הלשון העברית, בעריכת ' בלאו, ירושלים תשנ"ו, עמ' 7–13
- H. Blanc, "Dialect Research in Israel", *Orbis* 5 (1956), pp. = 1956 בלנק, 185–190
- H. Blanc, "A Note on Israeli Hebrew 'Psycho-Phonetics'", = 1956 בלנק, *Word* 12 (1956), pp. 106–113
- בן-טולילה, תשמ"ט = ' בן-טולילה, קורפוס הצרפתית של מונטריאול: דגם אפשרי לחקר העברית המדוברת, בלשנות עברית 27 (תשמ"ט), עמ' 13–28
- Y. Berglund, "Exploiting a Large Spoken Corpus: An = 1999 ברגלונד, End-User's Way to the BNC", *International Journal of Corpus Linguistics* 4 (1999), pp. 29–52
- R. A. Berman, "Modern Hebrew", *The Semitic Languages*, = 1997 ברמן, ed. R. Hetzron, London 1997, pp. 312–333
- L. Glinert, *The Grammar of Modern Hebrew*, Cambridge = 1989 גלינרט, 1989
- M. S. Devens, "Oriental Israeli Hebrew: A Study in = 1980 דבנס, Phonetics", *Afroasiatic Linguistics* 4 (1980), pp. 127–141; 7 (1980), pp. 26–37
- M. S. Devens, "Misconceptions about Accent and National = 1981 דבנס, Origin among Native Israeli Hebrew Speakers: A Preliminary Report", *Hebrew Annual Review* 5 (1981), pp. 21–36

ָרום, בדפוס = ר' וָרום, "הערות מתודולוגיות על כינון מאגר העברית המדוברת בישראל", מדברים עברית: לחקר הלשון המדוברת והשונות הלשונית בישראל, בעריכת ש' יזרעאל, תעודה, יח (בדפוס)

Sh. Izre'el, B. Hary and G. Rahav, "Designing = 2002, הרי ורהב, "CoSIH: The Corpus of Spoken Israeli Hebrew", *International Journal of Corpus Linguistics* 6 (2002), pp. 1–27

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

T. McEnery and A. Wilson, *Corpus Linguistics = 1996* (Edinburgh Textbooks in Empirical Linguistics), Edinburgh 1996

M. McCarthy, *Spoken Language and Applied Linguistics*, = 1998, Cambridge 1998

קדרי, תשמ"ד = מ"צ קדרי, "מבוא: מצב המחקר בעברית הישראלית", רשימת ספרים, מאמרים ועבודות דוקטור על העברית של ימינו שנכתבו עברית וראו אור בישראל בשנים תש"ח-תש"ס, בעריכת ב"צ פישלר, ירושלים תשמ"ד, עמ' 16-1

קדרי, תשמ"ח = מ"צ קדרי, "על חקר העברית המודרנית", דברי הקונגרס העולמי התשיעי למדעי היהדות, חטיבה ד, כרך א, ירושלים תשמ"ח, עמ' 85-92

קדרי, תשנ"ו = מ"צ קדרי, "דחיפותו של סקר על העברית הספרותית החיה", הלשון העברית בהתפתחותה ובהתחדשותה: הרצאות לרגל מלאות מאה שנה לייסוד ועד הלשון העברית, בעריכת י' בלאו, ירושלים תשנ"ו, עמ' 127-147

G. Kennedy, *An Introduction to Corpus Linguistics*, Studies in Language and Linguistics, London 1998

S. Crowdy, "Spoken Corpus Design", *Literary and Linguistic Computing* 8 (1993), pp. 259–265

הפניות לאתרי אינטרנט

- אתר של נתונים על מאגרי לשון:

<http://www.ruf.rice.edu/~barlow/corpus.html>

- דפים מאתר המאגר הלאומי הבריטי:

<http://info.ox.ac.uk/bnc/what/balance.html>

http://info/ox/ac/uk/bnc/what/spok_design.html

- אתר מעמ"ד:

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