המודל התכנוני של מאגר העברית המדוברת בישראל (מעמ"ד)

בנימין הרי ושלמה יזרעאל

מטרות כינון המאגר

1. יצירת מאגר של העברית הישראלית המדוברת כתשתית למחקר שיטתי. מחקר המאגר יקיף מיגוון רחב של נושאים הקשורים בשפה העברית ובמתודולוגיה הכללית של חקר הלשון המתבסס על מאגרי לשון.
2. הפצת המאגר לציבור במולטימדיה ובדפוס. ההפצה באמצעים אלקטרוניים — DVD-ROM, CD-ROM והאינטרנט — תיעשה כך שההקלטות ותמלילים יוצגו במקביל ובשילוב דרכי תיעוד וניתוח נוספות.

* מאמר זה הוא סיכום תמונת מצב המחקר לקראת כינון מעמ"ד כפי שהיה בפני צוות המאגר בפברואר 2000, ואשר אליו התייחסו גיורא רהב ורגינה וֶרוּם בהרצאותיהם המובאות בכרך זה. לפירוט רב יותר של הדברים, למצע התיאורטי שעליו נשען מודל מעמ"ד ולעדכונים, אנו מפנים את המתעניינים למאמרים אלה:

שלמה יזרעאל, בנימין הרי וגיורא רהב. תשס"ב. לקראת כינון מאגר העברית המדוברת בישראל. **לשוננו** ס"ד: 265–287.

Shlomo Izre'el, Benjamin Hary and Giora Rahav. 2001. Designing CoSIH: The Corpus of Spoken Israeli Hebrew. *International Journal of Corpus Linguistics* 6/2: 171–197.

צוות החוקרים של **מאגר העברית המדוברת בישראל**: שלמה יזרעאל (ראש התוכנית וחוקר ראשי); בנימין הרי (חוקר ראשי); ג'ין די בואה (אנליסט המאגר); מירה אריאל (חקר השיח ופרגמטיקה); גיורא רהב (סוציולוגיה וסטטיסטיקה). צוות יועצים: אליעזר בן רפאל (סוציולינגוויסטיקה — היבטים סוציולוגיים); יעקב בן טולילה (סוציולינגוויסטיקה — היבטים בלשניים); אוטו יסטרו (תיעתוק, פונולוגיה ודיאלקטולוגיה); שמואל בולוצקי (פונולוגיה, מורפולוגיה); ג'פרי קאן (תחביר); אילנה שוהמי (חינוך לשוני).

מאפייני מעמ"ד

* הקלטות דיגיטליות בקלטות שמע
* מבחר הקלטות דיגיטליות בווידיאו
* תמלילי כל הטקסטים במאגר בכתיב עברי
* תעתיק פונטי של קטעים נבחרים
* גלוסות של קטעים נבחרים
* תרגום לאנגלית של קטעים נבחרים

מהות המאגר

מאגר העברית המדוברת בישראל (מעמ"ד) ייצג את מיגוון השונות של העברית המדוברת בישראל היום. בכוונתנו לכלול מידגם מייצג של חלופות דימוגרפיות ותלויות הקשר. החלופות הדימוגרפיות מזוהות עם קבוצות דוברים שונות: גיאוגרפיות, אתניות, סוציואקונומיות, וחברתיות (גיל, מין, חינוך, מקצוע, נטייה מינית, וכו'). חלופות תלויות הקשר הינן פונקציה של מצבים שונים, כגון שיחה (פנים אל פנים, טלפונית) או נאום (ספונטני, מתוכנן, מצוטט מן הכתב).

בכינון המאגר הבאנו בחשבון את המבנה המיוחד של החברה הישראלית. בקרב האוכלוסייה הדוברת עברית מספר הדוברים הילידיים והלא ילידיים שווה. עובדה זאת, יחד עם המורכבות שנוצרה עקב ההיסטוריה המיוחדת של העברית החדשה, היא בעלת חשיבות מכרעת לגבי השאלה של הרכב הטקסטים שייכללו במאגר. אנו החלטנו לכלול טקסטים של דוברים ילידיים ושל דוברים לא ילידיים כאחד. לישראל אישים בעלי השפעה שאינם נמנים עם הדוברים הילידיים. בנוסף לכך, האוכלוסייה הולכת ומתרחבת תדיר בעקבות הזרימה המתמדת של עולים לארץ ישראל. אף זה גורם בעל השפעה על תנודות במערכת הלשונית, שחשוב לתעד. המצב הוא כזה, שתיעוד לשונם של הדוברים הילידיים בלבד לא ישקף את העברית בת זמננו, וודאי לא את המצב החברתי־לשוני בישראל על כל מורכבותו.

שאיפתנו היא ליצור מאגר שייצג נאמנה את כלל סוגי הדיבור העברי בישראל, ויכלול מיגוון רחב ככל האפשר של דוברים ושל הקשרי שיח. השם שניתן למפעל, **מאגר העברית המדוברת בישראל**, משקף שאיפה זו.

ממדי המאגר

סה"כ 5,000,000 מילה;
1000 תאים (=יחידות) המכילים 5000 מילה כל אחד;
5% מהמאגר יוקלט בווידיאו.

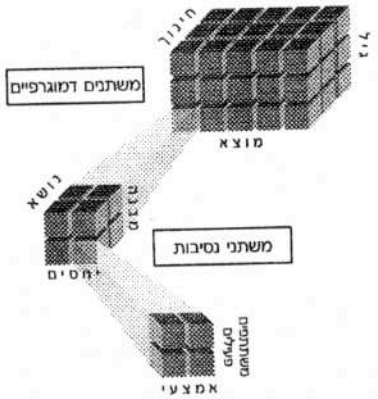
המידגם המייצג

על מאגר לשון מייצג להציג שני סוגי חלופות לשוניות: חלופות דימוגרפיות וחלופות תלויות הקשר. אי־לכך, עלינו לאסוף נתונים על פי שני סוגי קריטריונים:

* קריטריונים דימוגרפיים
* קריטריונים לבחינת הקשרי השיח

מעמ"ד יורכב משני תת־מאגרים: מאגר ראשי ומאגר משלים. המאגר הראשי יהווה את חלק הארי של מעמ"ד, ויכלול כ־95% מכלל הנתונים. המאגר המשלים יכלול כ־5% מהנתונים.

המאגר הראשי יורכב מתאים שייבנו על פי קריטריונים דימוגרפיים משולבים בקריטריונים לבחינת הקשרי השיח. נשתמש בתבנית שציריה העיקריים יהיו ציר דימוגרפי וציר הקשרי השיח. כל אחד משני הצירים הוא עצמו מהווה מרחב רב־ממדי: כל אחד מ־45 (=3×3×5) התאים בתבנית הדימוגרפית (ר' בהמשך) מכיל 8 (=2×2×2) תאים שונים על פי דגימת ההקשר. כל תא מתאים אלה עשוי להכיל בנוסף 4 (=2×2) תאים.



הנתונים במאגר המשלים ייאספו על פי קריטריונים של הקשרי שיח, תוך התחשבות במאפיינים דימוגרפיים. 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 ילידיים ממוצא צפון־אפריקאי, צעירים ובעלי השכלה על־תיכונית, יוכל לקבל את המידע על פי הקריטריונים האלה.

מחובתנו להדגיש כאן את ההבחנה בין הקריטריונים שישמשו לדגימה לבין הקריטריונים שישמשו לניתוח החומר הלשוני. האתגר העיקרי בתיעוד מידגמי של אוכלוסייה המונה מעל ארבעה מיליון נפש (מעמ"ד יכיל נתונים מדיבוריהם של תושבי ישראל מגיל 16) 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 הדימוגרפיים כשלעצמם נערכים על פי רובד נוסף, הוא רובד הקשרי השיח על מאפייניו השונים. בעוד המאגר יכיל בבסיסו נתונים טקסטואליים שייאספו על פי בדיקה מידגמית אקראית של שכבות האוכלוסייה, מידע דימוגרפי מפורט ומידע לגבי הקשרי השיח שיילקט ויאוחסן תוך כדי ביצוע ההקלטות, יאפשר עיבוד אנליטי סוציולינגוויסטי ולשוני מורכב.

דגימת האוכלוסין עצמה תיעשה על פי קריטריונים סטטיסטיים כדי לשקף כמותית את כלל האוכלוסייה. לעומת זאת, איסוף נתונים המתבצע על פי קריטריונים אנליטיים לא בהכרח יהא מייצג מבחינה סטטיסטית. במילים אחרות, המשתמש בבסיס הנתונים של המאגר יוכל לשלוף נתונים לשוניים מתוך בסיס נתונים, שאמנם יהיה ערוך על פי קריטריונים סוציולינגוויסטיים, אולם הנתונים שיוצאו לא בכל מקרה יהוו ייצוג כמותי של האוכלוסייה המיוצגת על ידי טקסטים אלה. מעמ"ד ישאף לגשר בין אינספור חלופות השיח הנהוגות בקרב קהילת הדוברים של העברית הישראלית לבין ייצוגיותה במאגר על ידי אפיון השונות במונחים דימוגרפיים ותלויי הקשר, בהתאם לקריטריונים סטטיסטיים ואנליטיים.

הקריטריונים האנליטיים לגבי הנתונים הסוציולינגוויסטיים במאגר יהיו שונים ורבים יותר מאשר אלה שישמשו אותנו בעת כינון המאגר ואירגונו. כדי להקל על המשתמש ולאפשר הסקת מסקנות הנשענות על בסיס איתן, ובלא להפר את עקרון הייצוגיות, יהא עלינו להגדיר כמה קריטריונים אנליטיים, הן דימוגרפיים והן בלשניים, עוד בשלב אירגון החומר.

המאגר הראשי

קריטריונים אנליטיים

המאגר יכלול 5,000,000 מילה: 1,000 תאים בני 5,000 מילה כל אחד. המאגר יורכב איפוא מ־1,000 תאים של טקסט מוקלט. "תא" הוא היחידה הסוציולינגוויסטית הבסיסית במעמ"ד. התא הוא קטע דיבור מוקלט בן 5,000 מילה של טקסט רציף ומלוכד או כמה קטעים שיחדיו כוללים 5,000 מילה. המאגר הראשי, שיהווה 95% ממעמ"ד, יכיל 950 תאים. 900 תאים מתוך אלה יידגמו בדגימה אקראית. מלבדם ייאספו נתונים בדגימה לא מייצגת לעוד 50 תאים כדי לתעד את דיבורן של קבוצות מיוחדות, שאנחנו מניחים שהן משמעותיות למצב הלשוני בישראל (כגון אנשי צבא, חרדים, ישראלים ששהו תקופות ממושכות בחו"ל וכיו"ב). המאגר המייצג האקראי, שהוא עיקר המאגר הראשי, יכיל איפוא 900 תאים שווים בגודלם.

א. קטגוריות דימוגרפיות

הגישה שננקטה על ידינו לכינון המאגר היא גישה תלויית 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 אם לעת עתה אין אנו יכולים אלא להעלות השערות באשר לדומה ולמפריד ביניהן). מובן מאליו כי השונות מעוגנת במאפיינים דימוגרפיים ועלינו לנתחם כראוי. הנחת העבודה שלנו מתבססת על שלושה קריטריונים דימוגרפיים שנראים לנו החשובים ביותר ביצירת השונות הלשונית בישראל: (1) עדה או דת, מקום לידה וארץ מוצא המשפחה; (2) גיל; (3) השכלה.

(1) מקום לידה, ארץ מוצא המשפחה, ועדה או דת (5 קטגוריות):

1. יהודים, ילידי הארץ, בני אב מאסיה־אפריקה
2. יהודים, ילידי הארץ, אחרים
3. יהודים, ילידי חו"ל, שנת עלייה עד 1965
4. יהודים, ילידי חו"ל, שנת עלייה אחרי 1965
5. לא יהודים (מוסלמים, נוצרים, דרוזים)

(2) גיל (3 קטגוריות):

1. צעירים
1א. בני 16–18
1ב. בני 19–27
2. בוגרים (28–50)
3. מבוגרים (בני 51 ומעלה)

(3) השכלה (3 קטגוריות)

1. מי שלא סיים תיכון
2. מסיימי תיכון (או שנת לימודים אחת נוספת)
3. בעלי השכלה על־תיכונית

הציר הדימוגרפי של תבנית התאים הרב־ממדית שתשמש כבסיס לניתוח מעמ"ד הוא עצמו רב־ממדי, איפוא, וכולל שילוב הקטגוריות של משתני מקום הלידה, המוצא והעדה או דת עם משתני הגיל ועם משתני השכלה. קריטריון משמעותי 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 ונשים בכל קטגוריה אנליטית תהיה בהתאם לחלוקת המינים באוכלוסייה, דהיינו 1:1.

ב. קטגוריות הקשרי השיח

כדוגמת הציר הדימוגרפי, גם ציר הקשרי השיח מעוצב כתבנית רב־ממדית. בעת קביעת החלופות תלויות ההקשר לקחנו בחשבון חמישה משתנים, שלושה עיקריים ושלושה משניים.

משתנים עיקריים:

(א) יחסים בינאישיים: קירבה מול ריחוק (+/–קירבה)

במשתנה (א) משתקפים היחסים האישיים. כשיש קירבה בין בני השיח, כלומר, כאשר בני השיח הם בני משפחה או חברים, נציין + קירבה.

(ב) מבנה השיח: תלוי־תפקיד מול שוויון (+/–תפקיד)

במשתנה (ב) משתקף מבנה השיח. כאשר מובחן מבנה שיח שבו חלוקת תפקידים ברורה בין בני השיח (למשל, כשיש תפקיד סמכותי למי מהם), נציין + תפקיד.

(ג) נושא השיח: אישי מול לא אישי (+/–אישי)

שלוש הקטגוריות שנמנו (א–ג) יבואו לידי ביטוי במאגר בכל 8 הצירופים האפשריים שלהם (סך הכול 2^3 נסיבות). לעומת זאת, המשתנים המשניים דלהלן (ד–ה) יישמו רק בחלק מהתבנית, בהיותם הרבה פחות שכיחים ברוב קהיליות 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) אמצעי: טלפון מול פנים־אל־פנים (+/–טלפון)

גם שיחה באמצעות הטלפון תובחן כמהותית רק במקרים שיאופיינו כך.

כדי להמחיש את המאפיינים הבינריים הללו והשתלבותם, הנה דוגמאות אחדות:

(1) שיחת יומיום בין בני משפחה (+קירבה; –תפקיד; +אישי; –מונולוג; –טלפון)
(2) שיחה על פוליטיקה בין חברים קרובים (+קירבה; –תפקיד; –אישי; –מונולוג; –טלפון)
(3) שיחת יומיום במשפחה בעלת גינונים מסורתיים (+קירבה; +תפקיד; +אישי; –מונולוג; –טלפון)
(4) ראיון עבודה טלפוני (–קירבה; +תפקיד; –אישי; –מונולוג; +טלפון)
(5) נאום פוליטי (–קירבה; +תפקיד; –אישי; +מונולוג; –טלפון)

דגימה

(א) דגימה דימוגרפית

מטרת כינון תאים מייצגים מבחינה סטטיסטית היא לאסוף נתונים מספיקים למחקר סוציולינגוויסטי או בלשני. נשתדל לכונן מאגר שגם יהווה תמונת המערכת הלשונית בכללה, וגם ייצג נאמנה מיגוון רחב של שונות לשונית. ביצירת מידגם אוכלוסין לצורך זה או אחר אפשר לחלק את האוכלוסייה

לתת־קבוצות ולדגום כל אחת מהן בנפרד ("ריבוד" הדגימה). ריבוד דגימה עשוי להידרש מכמה טעמים, שהחשובים בהם:

1. לתת־קבוצות מסוימות מאפיינים השונים משמעותית מאלה של כלל האוכלוסייה (בעיקר אם הקבוצה היא קטנה יחסית או אינה מעורה דייה באוכלוסייה);

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

3. הקבוצה קטנה וברצוננו לתת לה ייצוג מוגבר.

כינון מידגם ללא ריבוד אין פירושו בהכרח שתוצר הדגימה לא יהיה מייצג דיו, או שיש בו פחות ערך מאשר מידגם מרובד. העדר הריבוד משמעו שאנו מניחים לתהליכים הסטטיסטיים האקראיים לייצג את האוכלוסייה הנדגמת במידה סבירה. לפיכך, בשלב זה, דגימה אקראית של המאגר לפי אזורי מגורים תשמש מידגם ראשוני בבחירת האינפורמנטים.

חלקים נרחבים באוכלוסייה אכן יתועדו בכמות ניכרת, ואנו צופים שיצטבר די חומר לייצוג לשוני הולם. עם זאת, רמת ייצוגן של קבוצות אוכלוסין מסוימות לא תספיק לשם הסקת מסקנות כלליות על מאפייני לשונה של קבוצות אלו. אף על פי שמעמ"ד יכלול ייצוג מסוים של תת־קבוצה כזאת, התא או התאים המתועדים יוכלו רק לכוון את החוקרים באשר לטיפול הדרוש למחקר מקיף מן הסוג הזה ולא לייצגם נאמנה. במחקרים מן הסוג הזה יש לאסוף נתונים מאוכלוסיית היעד בנפרד. החוקרים יוכלו להשתמש במעמ"ד כמקור להשוואה, כמסגרת התייחסות לגבי כלל האוכלוסייה, או, מה שחשוב יותר, כדי לקבל מידע ראשוני באשר לסוג המחקר הנחוץ לכל אחת מאוכלוסיות היעד או קהילות הדוברים המסוימות האלה. לגבי הקבוצות הלשוניות הגדולות, המאגר בן חמישה מיליון המילה יספיק לעריכת מחקרים מקיפים על מיגוון רחב של היבטים בלשניים וסוציולינגוויסטיים של העברית המדוברת בישראל.

(ב) דגימת הקשרי השיח

עריכת מידגם מייצג על פי נתונים דימוגרפיים היא טכניקה שכיחה הנהוגה בדגימת אוכלוסין, ומתבצעת על ידי דגימה אקראית של האוכלוסייה. לעומת זאת, הניסיון שנצבר בבניית מאגר לשון שייצג נאמנה גם את הקשרי השיח מועט ביותר.

כדי להבטיח ייצוג אמין של השונות הלשונית, הדימוגרפית ותלויית ההקשר כאחת, יישמר עקרון האקראיות בעת דגימת כל הנתונים הטקסטואליים. דגימת

הקשרי השיח תתבצע לאחר איסוף כל ההקלטות של האוכלוסייה הנדגמת דגימה דימוגרפית. כל אלה שעלו במידגם האוכלוסין האקראי יתבקשו להקליט את כל מהלכיהם והקורות אותם במשך שלושה ימים רצופים. כל הקלטה כזאת תיבדק, ויוסרו מירווחי שתיקה וקטעי דיבור לא ברורים. מכל החומר הנקי ייבחר באקראי קטע של טקסט באורך של שעה, והוא שיהווה את מקור הנתונים שייכלל במאגר הראשי. בחירה אקראית זו עשויה לשקף את ביזור מצבי השיח הטבעיים, הן על פי היקרויותיהם בזמנים שונים, הן בהתאם להקשרי השיח, והן בהתאם לסוגי השיח המזומנים לסוגי אוכלוסייה שונים.

גישור הפער בין תביעות הסטטיסטיקה לבין הקריטריונים האנליטיים: היערכות התבנית הרב־ממדית

התוכנית הבסיסית של מעמ"ד מקנה 20 תאים לכל חלופה דימוגרפית. היות שיש 45 חלופות דימוגרפיות, משילובן יצטברו 900 תאים. לא כל החלופות תואמות זו את זו. למשל, קריטריון מקום הלידה 3 (יהודי שאינו יליד הארץ ועלה לפני 1965) אינו תואם את קבוצת הגיל הראשונה (16–27 שנים) ואף לא את כל מרחב קבוצת הגיל השנייה (28–50 שנים), שגם בה ילידי התקופה שאחרי 1965. בנוסף על כך, חלק מהנדגמים בתת־קבוצת הגיל 1א (בני 16–18) אינם יכולים להיות בעלי השכלה על־תיכונית (משתנה החינוך מס׳ 3). אי־לכך, כבר על פי התכנון הראשוני לא תוכל התבנית להתמלא בכל 900 תאיה. כמו כן אנו סבורים שלא כל החלופות האפשריות מבין החלופות תלויות ההקשר זמינות וקיימות עבור כל אחת מן החלופות הדימוגרפיות.

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.

אירגון החומר הטקסטואלי שנאסף הוא השלב הבא. עריכת הנתונים לשימוש מחקרי תתבסס על העקרונות שנקבעו בכינון תבנית התאים ותתייחס לקריטריונים הדימוגרפיים ולקריטריונים תלויי ההקשר ששוקללו כפי שתואר לעיל. יעמדו לרשותנו טקסטים שנבחרו באקראי מתוך מצאי טקסטים שהופקו על ידי אנשים שנבחרו באקראי במיגוון רחב של הקשרי שיח. הטקסטים הללו ייערכו בתבנית הרב־ממדית, כשכל יחידת טקסט תוכנס אל התא התואם לה על פי הנתונים הדימוגרפיים של דובריה ועל פי הקשר השיח המתאים. חלוקת החומר הטקסטואלי לתאים תיעשה על פי עקרון האקראיות ובכימות הולם.

דהיינו, תבנית התאים תתמלא במצאי הטקסטים שהבחירה האקראית העמידה לרשותנו ולא על פי מספר התאים שהעיצוב העקרוני העמיד לרשות כל יחידה דימוגרפית. גישה זו תאפשר לנו לעמוד על מיגוון דגמי השיח הקיים בקהילת הדוברים וללמוד אודות הדגמים כמותית ואיכותית, כפי שיעלה משילוב הקריטריונים הדימוגרפיים והקריטריונים תלויי ההקשר. במילים אחרות, המיון לתאים יניב ניתוח ראשוני של הקשרי השיח המאפיינים חלקי אוכלוסייה שונים, וכן ניתוח ראשוני של היקף השימוש של החלופות הלשוניות הללו ביחס לחלופות אחרות. כפי שצוין לעיל, כל הנתונים הסוציולינגוויסטיים יהיו נגישים לכל דורש, וחוקר המאגר יוכל לשלוף טקסטים לצרכיו ועל פי נתונים סוציולינגוויסטיים כחפצו. עם זאת, כדי להפיק תוצאות משמעותיות בניתוח לשוני, אנו מציעים את המשתנים שפורטו לעיל כבסיס לשליפת נתונים.

תאים נספחים

צירופים מסוימים, דימוגרפיים או דימוגרפיים ותלויי הקשר, 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 הדימוגרפי של קהילת הדוברים בישראל, או של מצאי הקשרי השיח הטבעיים לקבוצות השונות. בכל זאת, במקרים מסוימים יידרשו תיקונים, אם מפאת פגם כלשהו במידגם ואם מתוך שיקול למתן ייצוג יתר לקבוצה מסוימת. ייצוג יתר יישקל עבור קבוצה, או קבוצות, שנראה כי יש לה השפעה מיוחדת על ההתנהגות הלשונית של דוברי העברית בישראל. עבור קבוצות אלה ייכונו תת־מאגרים לא מאוזנים, והם שיהוו את מיכסת 50 התאים הנוספת על 900 תאי המאגר הראשי שהוזנו באופן מאוזן. כפי שניתן לנבא בשלב זה, התאים האלה יכילו נתונים מדיבורם של אנשי צבא, חרדים, או דוברים ילידיים של עברית ששהו תקופות ארוכות מחוץ לישראל.

לשון הצבא תזכה לתשומת לב מיוחדת. השירות הצבאי בישראל הוא חובה ונמשך שלוש שנים לגברים ושנה ותשעה חודשים לנשים. לאחר מכן, גברים משרתים עוד במשך תקופה מסוימת בשירות מילואים, לעתים עד גיל 49. רבים אחרים משרתים בצבא הקבע או בכוחות הביטחון. בהיות מדינת ישראל מדינה קולטת עלייה במהותה, השירות הצבאי שימש מאז ומתמיד ככור ההיתוך של החברה הישראלית. על שום חשיבותו הסגולית של הצבא בחיי החברה בישראל, השפעתו על העברית עצומה. בעיקר ניכרת ההשפעה הזו באוצר המילים ובמטבעות לשון, אך אין ספק שתימצא גם מעבר לתחומים אלה. ודאי

שהמאגר הראשי יכיל בתוכו גם תת־מאגר של הקלטות מהצבא. בשלב זה אין לדעת אם תת־מאגר זה ייווצר במידגם האקראי, או שייערך בנפרד.

המאגר המשלים

המאגר הראשי צפוי לתת ייצוג הולם לרוב סוגי הדיבור העיקריים. עם זאת, התבנית המתוארת לעיל אינה מייצגת תחומים אחדים של לשון הדיבור שחשיבותם רבה כדי כך שמן הראוי שיימצא מקומם במעמ"ד. כך הוא דיבורם של חברי הכנסת בבית הנבחרים, הדיבור בבתי משפט, ומעל לכול לשון אמצעי התקשורת (טלוויזיה ורדיו). בהיות כלל האוכלוסייה חשופה למקורות אלה במידה זו או אחרת, חלופות לשוניות אלה, הגם שאינן משקפות את דיבורם של דוברים רבים, 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 המיוצגות במאגר המשלים מפורטות להלן. ייערך מעקב מתמיד אחר תוצר איסוף התאים הזה כדי לבדוק את הצורך להשלימו על ידי תוספת תאים מקבילים מפי דוברים בעלי רקע דימוגרפי שונה במובהק.

			ספונטני	לא־ספונטני	
				דיבור חופשי	מוקרא מן הכתב
טלוויזיה	1א	שידור רגיל	+		
	1ב	שידור רגיל		+	
	1ג	שידור רגיל			+
	2א	שידור ספורט	+		
	2ב	שידור ספורט		+	
	2ג	שידור ספורט			+
	3	ראיון	+	+	
	4	רב־שיח	+	+	
	5	סרט			+
	6	פרסומת			+
רדיו	7א	שידור רגיל	+		
	7ב	שידור רגיל		+	

	ג7	שידור רגיל		+
	א8	שידור ספורט	+	
	ב8	שידור ספורט		+
	ג8	שידור ספורט		+
	9	ראיון	+	+
	10	רב־שיח	+	+
	11	שיחות טלפון	+	+
	12	פרסומת		
כנסת	א13	נאום		+
	ב13	נאום		+
	ג13	מונולוגים; דיאלוגים	+	
בית משפט	א14	נאום		
	ב14	נאום		+
	ג14	מונולוגים; דיאלוגים	+	

ובכן, המאגר המשלים, הבנוי בבסיסו על קטגוריות של הקשרי שיח, כולל 26 תאים בני 5,000 מילה כל אחד, שהם 130,000 אלף מילה, כ־2.6% מן המאגר כולו. משום שוודאי נצטרך להוסיף למאגר המשלים תאים נוספים על פי נתונים דימוגרפיים שונים של הדוברים, הערכתנו היא כי המאגר המשלים יוכפל בגודלו, ובסופו של דבר יכלול כ־5% מן המאגר כולו.