דגימת אוכלוסייה לכינון מאגר מייצג

גיורא רהב

אין בכוונתי לתאר את מה שנעשה עד כה, את הישגינו או את העבודה הרבה שהשקענו, כי אם לדון בתוכניות, ברעיונות ובתקוות. נדבר על העתיד, על מה שברצוננו לעשות.

אנו מנסים כעת למצוא דרך לייצג בה כראוי את העברית המדוברת כפי שמשתמשים בה כיום דובריה הטבעיים. מובן שהדרך הטובה ביותר לעשות זאת היא להקליט עכשיו את כל מה שאומרים כולם, ובשנה הבאה להעמיד את ההקלטות לרשות הכול. ברור גם כי אין זה אפשרי. לכן, הדרך היחידה לעשות זאת תהיה להתבסס על מידגם. כלומר, להתמקד בתת־קבוצה של דוברים ושל דיבורים שתייצג את האוכלוסייה כולה, על דפוסי הדיבור שלה. השאלה היא: כיצד נוכל להבטיח שקבוצה זאת אכן תהיה מייצגת?

גישה אחת לבעיית הדגימה היא לערוך דגימה של "דוברים טיפוסיים", המשתמשים ב"לשון טיפוסית". הבעיה בפתרון זה היא ש"דוברים טיפוסיים" הם מאוד לא־טיפוסיים (או לפחות — נדירים). מה שנחשב בעינינו כ"דובר טיפוסי" או כ"דיבור טיפוסי" עלול לשקף את התפישות והסטריאוטיפים שלנו יותר מאשר את המציאות של השפה ושל דובריה. לכן נחוצה לנו דרך אחרת לייצג את האוכלוסייה בדרך שקולה ושיטתית. ניתן לעשות זאת על ידי הכנת דגימה סטטיסטית של דוברים ושל הדיבור שלהם.

לשם כך, בראש ובראשונה עלינו לקבוע לעצמנו גבולות ולתחום בהם את כל מה שברצוננו לייצג. אם כן, אנסה להגביל את עצמי לשפה המדוברת הטבעית, לשפה הדבורה.[1] כלומר, לא אטפל ב"לשון מעוצבת", הלשון שאדם נוקט כדי שאדם אחר או צד שלישי (קהל צופים או מאזינים) ישמע אותה. אתייחס רק,

* הטקסט המתפרסם כאן הוא תרגום ערוך של תמליל ההרצאה שניתנה בכנס אטלנטה.

1 למינוח ר׳ בדברי הפתיחה, עמ׳ טו–טז לעיל.

או בעיקר, לשפה כפי שהיא מדוברת בשיחות אותנטיות, בין בני שיח שדבריהם לא נקבעו ולא הוכנו מראש. משמעות הדבר היא כי לשון התיאטרון או לשון הדוברים בטלוויזיה או ברדיו לא תיכלל בדיון זה. מלבד זאת, מאחר שנדגום בעיקר שיחות, מספר האנשים (הדוברים המוקלטים) המעורבים בשיחות יהיה גדול בהרבה ממספרם של מקליטי הדיבור.

ובכל זאת, הבעיה העיקרית היא לדגום את הדוברים המוקלטים ואת דיבורם כך שנהיה בטוחים כי יש בידינו ייצוג הולם לקהילה דוברת העברית כולה ולדיבור העברי בישראל כולו. משמע, שעלינו לדגום קהילות ותת־קהילות, לדגום תת־תרבויות ולוודא שלפחות תת־התרבויות המשמעותיות ביותר ייוצגו במידגם כראוי.

שנית, עלינו לדגום סיטואציות של דיבור והקשרים של דיבור. יש להקדיש לעניין מחשבה ותכנון, משום שאיננו רוצים שהמאגר ייצג רק סוג אחד של מצבים, כגון שני אנשים מדברים, האחד מראיין והשני משיב על שאלותיו. ברצוננו לייצג את כל טווח הנסיבות המזמנות שימוש בלשון.

ולבסוף, אולי כתת־קטגוריה של דגימת סיטואציות, יש לדגום תקופות ופרקי זמן. כוונתי לכך שנסיבות הדיבור משתנות בקביעות מסוימת במהלך היום, השבוע או השנה. אופי הדיבור בתקופות של עומס בעבודה או בזמן חופשה עשוי להיות שונה מדיבור באמצע היום, כאשר כולם עסוקים בעבודה שגרתית. אופני דיבור המשתנים בסדירות יהיו שונים מאלה המשתנים ללא סדירות, כגון "אופנות" בדיבור. כך שהשאלה הניצבת בפנינו היא כיצד לדגום את כל אלה.[2]

לשם עיצוב הדגימה, נוכל לנקוט שתי גישות עיקריות, שונות לחלוטין זו מזו. האחת היא הגישה הלא־סטטיסטית, או "מידגם מכוון". כך נוכל אנו, האחראים לבחירת המידגם, להחליט מי ישתתף במידגם. לדוגמה, ניתן להחליט שייכללו בו עשרים דוברים ממוצא פולני, עשרה ממוצא ערבי, שני אתיופים וכן הלאה. זוהי דרך אחת.

דרך נוספת ליצור מידגם "נבחר", לא־סטטיסטי, היא לתת למידגם להתפתח מעצמו. לדוגמה, אם ידועים לנו מספר אנשים שמתאימים למידגם מבחינת שליטתם בשפה, או אם יש תת־אוכלוסייה שברצוני שתיוצג במידגם הסופי שלי, כל שעליי לעשות הוא לאתר ולזהות אותם. ברגע שאמצא אחד מהם,

2 חשוב לציין שהצגת הדברים כאן מפשטת מאוד את הבעיה. בפועל, עשוי להיות הבדל ניכר בין דגימת דוברים, דגימת דיבור, דגימת נסיבות ודגימת תקופות. על מנת לפשט ולהדגיש את ייצוגן של הקהילות השונות של דוברי עברית, יתמקד הדיון בדגימת הדוברים.

אראיין אותו ואקליט אותו, ואוכל לבקש ממנו לכוון אותי אל מספר אנשים נוספים הכלולים באותה קטגוריה, או השייכים לאותה תת־תרבות. גישה זו ידועה בשם "דגימת כדור־שלג". כמו הגישה הקודמת, גם שיטה זו תניב מידגם מכוון, לא־סטטיסטי, אך יש לה מספר יתרונות: היא מאפשרת שליטה טובה למדי במידגם, וניתן להוציא מידגם גדול למדי במשאבים מוגבלים מאוד. אני מעריך כי זוהי הסיבה שחלק ניכר מהעבודה שנעשתה בתחום התבססה על מידגמים שנבחרו בשיטות אלה.

מול שיטות הדגימה האלה, הפחות או יותר מכוונות, ישנה הגישה הסטטיסטית, המחייבת את שילובם של תהליכים אקראיים בבחירת המידגם. לדגימה הסטטיסטית כמה וכמה יתרונות. קודם כול, אם היא נעשית כהלכה, היא מבטיחה ייצוג הולם של האוכלוסייה כולה, האוכלוסייה כפי שנגדיר אותה מראש. זהו יתרון על־פני השיטות הקודמות, שלכל היותר מייצגות את מה שאנו מאמינים שהם המאפיינים העיקריים של האוכלוסייה. יתרונה של הגישה הסטטיסטית הוא בכך שהמידגמים המבוססים עליה אינם מושפעים מהשערותינו לגבי האוכלוסייה.

שנית, למרות שבדרך כלל ייצוגן של תת־אוכלוסיות שונות לא יהיה מדויק, השיטה מאפשרת לחשב את שיעור הטעות הצפוי. זהו יתרון משמעותי, כי הוא מאפשר לקבוע מהי מידת האמון שאפשר לרחוש לבסיס הנתונים. לבסוף, יתרון שלישי, אולי לא מדעי כל כך, אך בכל זאת חשוב: השיטה מקובלת על הקהילייה המדעית. עצם העובדה שגישה זאת לדגימה נחשבת לגישה התקנית בתחומי מדע שונים מביאה לכך שכשאומרים שהמידגם נבחר בשיטות סטטיסטיות, מתייחסים אליו ברצינות ומסקנות מנתוניו נראות הרבה יותר משכנעות. ייתכן בהחלט שבמקרים רבים זוהי הסיבה העיקרית להכרעה לטובת הגישה הזאת.

דגימה סטטיסטית מתבססת בעיקר על שני עקרונות חשובים. הראשון שבהם הוא כי לכל אדם, לכל יחידה באוכלוסייה, יש אותם סיכויים להיכלל במידגם. כך צריך להיות, לפחות בשלב ההתחלה, לפני שאנו מתחילים בהליכי הדגימה. משמעות הדבר היא שהדגימה חייבת להתבסס על תהליך אקראי, תהליך בו אין לחוקר כל אפשרות לקבוע את נבדקיו. זהו תהליך אקראי במובן זה שהוא אינו מושפע (ואינו מוטה) על ידי רצונות או אמונות כלשהם, או על ידי כל גורם אחר שיעלה בדעתנו.

עיקרון שני הוא כי המידגם חייב להיות רחב היקף. אין ביכולתי לומר מה ייחשב רחב היקף. איני מסוגל להעריך זאת. ברוב סקרי דעת הקהל בישראל

היקף המידגם נע בין 500 ל־1,500 איש, אך גם זה משתנה. הדבר תלוי במידה רבה במורכבות הנתונים, בניתוחים הנדרשים ובמיגוון (או בשונות) שלהם. לדוגמה, אם אנו יודעים מלכתחילה שנרצה להשוות בין דיבורם של גברים לזה של נשים, אנו כבר יודעים שנרצה לחלק את המידגם לשניים. אם נרצה להבדיל בין שתיים־שלוש קבוצות גיל עיקריות, הרי גם זה מחייב הבחנה בין מספר תת־קבוצות. אם אנו יודעים מראש שברצוננו לערוך תת־חלוקות כאלה, עלינו לוודא שכל תת־קבוצה שנרצה להתייחס אליה בנפרד תהיה מספיק גדולה כדי לייצגה. בהיותנו מוגבלים על ידי שיקולים מעשיים שונים, אנו מניחים שברוב המקרים מידגם של כ־1,000 איש עשוי להיות ראוי לייצג את רוב ההיבטים של הדיבור הישראלי.

את חובת הדגימה של עיתות הדיבור ואת הנסיבות השונות נוכל למנוע, אם ניתן יהיה לדגום את כולם, רק אם נצליח לשכנע כל משתתף או משתתפת במידגם לשאת איתו/איתה רשמקול במשך שבוע או שבועיים, או במשך חודש שלם, ולהקליט את כל מה שהוא/היא מעורב/ת בו. זה לא נראה לי מעשי. לדעתי, נוכל לעשות משהו דומה יותר לרעיון הבא:

בהנחה שנוכל לשכנע מידגם מספיק גדול של אנשים לשאת איתם רשמקול ולהקליט את עצמם במשך תקופות ארוכות (לא את הכול; לו דרשנו זאת, איש לא היה מסכים), הדבר יפתור את בעיית מצבי הדגימה. הסיבה לכך פשוטה: כל אדם נע בדרך כלל בין מיגוון של מצבים ותפקידים תוך כדי ביצוע פעילויותיו היומיומיות. מובן שאיש אינו מספיק לעבור את כל המצבים במשך יום אחד בלבד, ואפילו לא במשך חודש. עם זאת, אם נביא בחשבון את המידגם כולו על הקלטותיו המצטברות, הרי שהוא ינוע בטווח רחב, מייצג למדי, של נסיבות.

הצגת דברים זאת מציגה את בעיית הדגימה כולה בעיקר כבעיה של דגימת אנשים מתוך האוכלוסייה, כאשר אנשים אלה "ביאו" איתם מיגוון של סיטואציות, תפקידים ודפוסי דיבור. על מנת לקבל מידגם מייצג של אינפורמנטים, או של אנשים שייבחרו לביצוע ההקלטות, עומדות בפנינו שלוש אפשרויות עיקריות. ראשית, ניתן לדגום אנשים ממירשם האוכלוסין. דרך זאת מקבילה למצב שבו רושמים את שמותיהם של כל התושבים על פיסות נייר, מכניסים את כולן לכובע, מערבבים ומעלים בגורל את מספר השמות הדרוש. בפועל, הדגימה מתבצעת בטכניקה אחרת, אבל התוצאה היא אותה תוצאה. ההליך הטכני של בניית מידגם אקראי יכול להתבצע בעזרת מחשב, אבל חשוב לציין שאפשר לדגום אנשים מתוך מירשם האוכלוסין. אם אכן יידגם המירשם, הרי שבלי ספק נוכל לטעון שהמידגם מייצג את האוכלוסייה כולה (לפחות

המידגם הראשוני, לפני נשירה). מובן שייצוג זה אינו מושלם מסיבות שונות. לדוגמה, מספר אנשים עשויים לשהות בחו"ל, ואחרים לא יהיו רשומים משום שהם מהגרים לא חוקיים. עם זאת, על פניו יש למידגם כזה תוקף רב.

נוסף על כך, אם נוכל לגייס את שיתוף הפעולה של הלשכה המרכזית לסטטיסטיקה, שיש לה גישה ישירה לנתוני מיפקדי האוכלוסין, נוכל לקבוע מראש כמה מתכונות המידגם. לדוגמה, ניתן יהיה לבקש שייכללו במידגם כך וכך אנשים ממוצא מסוים ובגיל זה או אחר, המתגוררים באזור מסוים. שימוש בקריטריונים כאלה עשוי להבטיח כי מבנה המידגם יהיה זהה (בקריטריונים הללו) לזה של האוכלוסייה. לחלופין, ניתן להחליט כי לקבוצה מסוימת יהיה ייצוג יתר במידגם, כדי להבטיח את ייצוגה ההולם של קבוצה שעשויה להיות חשובה, למרות היותה קטנה יחסית.

לצד יתרונותיה הפוטנציאליים, יש לשיטה גם מספר חסרונות. החיסרון העיקרי הוא עלותה הגבוהה, שכן ברגע שיעלה שם כלשהו בגורל, יהיה צורך למצוא את האדם המסוים. אם הוא עבר דירה, יהיה עלינו לחפש אותו במישכנו החדש. אם הוא שוהה בחו"ל, ניאלץ להמתין לשובו ארצה, וכיוצא באלה.

גישה נוספת, פשוטה הרבה יותר, שהיא השיטה העיקרית המשמשת ברוב הסקרים, מבוססת בעיקר על דגימה אזורית של משקי בית, ומהם עולים השמות שייכללו במידגם. ניתן לחלק את הארץ כולה ליישובים וכל יישוב גדול יחולק לאזורים. פעולה זאת נעשתה כבר על ידי הלשכה המרכזית לסטטיסטיקה, שחילקה את היישובים הגדולים ל"אזורים סטטיסטיים" (בדומה ל־Census tracts במיפקד האוכלוסין בארה"ב). אפשר לדגום מכל הארץ מספר יישובים ואזורים סטטיסטיים, ובכל אזור כזה לדגום מספר נקודות. המראיינים ייגשו לדירה הראשונה בכל נקודה כזאת ויבחרו את אחד האנשים המתגוררים בדירה ההיא. לאחר מכן יִספרו, למשל, עשר דירות מהדירה הראשונה, וכך יגדירו את הדירה הבאה שידגמו. תוך כדי חזרה על הליך זה ניתן יהיה לדגום חמש, עשר או עשרים דירות בכל אחת מהנקודות. בכל דירה יבחר המראיין אדם אחד, לפי רשימה שנקבעה מראש.

שיטה זו טובה כמעט כמו הדגימה ממירשם האוכלוסין. מובן שאם ברצוננו להכתיב את מבנה המידגם, נוכל להקדים לראיון שאלון מיון קצר (לדוגמה, "בן כמה אתה?" או "היכן נולדת?" "היכן נולדו הוריך?" וכיוצא באלה) ולהחליט על בסיס שאלות אלה אם לכלול את האדם במידגם או לא.

לבסוף, שיטה נוספת המשלבת ראיון טלפוני עם הראיון שברצוננו לערוך, מתחילה בבחירה אקראית של מספרי טלפון מבין מספרי הטלפון הפרטיים

בארץ ובחיוג אליהם. נשאל מספר שאלות מיון, ועל בסיס שאלות אלה נחליט האם לכלול את המשיב במידגם או לא. אם נחליט שכן, נשאל לכתובתו, נקבע עימו פגישה ואז ננסה לשכנע אותו פנים אל פנים לשתף איתנו פעולה.

לא נותר לנו, אם כן, אלא לבחור באחת השיטות האלה, על סמך שיקולים טכניים וארגוניים שונים. הליכים אלה מותירים עדיין מספר בעיות, אך לא נוכל להתמודד עם כולן במסגרת הדיון בדגימה. ראשית, במקרים מסוימים עלולים להיווצר ליקויים טכניים שיפגעו במידגם ויטו אותו. ניתן לקוות שהבעיות הטכניות יתפזרו באקראי על פני כלל האוכלוסייה, כך שהן רק יצמצמו את המידגם מבלי להטות אותו. עם זאת, ראוי להבהיר שכל תהליך שמצמצם את המידגם עלול גם להטות אותו. כך, למשל, אם הנשירה (סירובים, קשיי איתור, הקלטות משובשות וכו') תגיע ל־5%, נוכל אולי לומר שהנשירה אינה מטה אותו. אך האם נוכל לומר זאת גם אם הנשירה תגיע ל־50%?, ל־80%? במקרים כאלה יהיה זה סביר להניח שיש הבדלים מסוימים בין המידגם שנבדק בפועל לבין האוכלוסייה, ויהיה חשוב לבדוק עד כמה מיוצגות לפחות הקבוצות החברתיות העיקריות שרצינו לכלול במידגם.

שנית, עלינו להגדיר את האוכלוסייה ולקבוע מי ייכלל בה. מה יהיו גבולות הגיל? האם ייכללו צעירים שגילם פחות מעשרים? אם כך, תהיה לנו בעיה לראיין את המשרתים בצבא. האם נכלול רק בני 20 ומעלה? 21 ומעלה? ואולי רק מבוגרים יותר? במקרה זה נחמיץ הרבה משפת המתבגרים. האם נכלול עולים חדשים שזה עתה הגיעו ארצה? הם עשויים לדעת רק מספר מילים בעברית, או אולי למדו עברית במשך שנים ארוכות והיא שגורה בפיהם. עלינו להחליט מספר החלטות בנוגע להכללתם במידגם, וכך גם אם לכלול זרים (כולל עובדים זרים ותיירים). יש מבקרים שנשארים כאלה שנים רבות, ויש גם סטודנטים זרים (במיוחד תלמידי ישיבות), השוהים בארץ שנים. כל אלה עשויים לשלוט היטב בעברית או לא לדעת אותה כלל, אך הם נמצאים כאן ועלינו להחליט החלטה מפורשת בדבר היכללותם או לא, ועל פי אילו קריטריונים.

סוג כזה של מחקר כרוך בבעיה טכנית לא קלה, הבעיה של חדירה לפרטיות. סוגיה זו עלולה להיות בעייתית במספר רמות. היא עלולה להגדיל את שיעור הסירובים הצפוי בקרב האוכלוסייה הנדגמת. ניתן להתגבר על הבעיה, אך עלינו לבדוק (במחקר טרומי, או אולי במספר מחקרים טרומיים) מה תהיה הגישה הטובה ביותר. בעיה נוספת עלולה להיווצר על ידי אנשים שתחילה יסכימו להשתתף, אך מאוחר יותר ינשרו מן המידגם. הדבר יכול להיעשות במוצהר, או בדיעבד, על ידי סגירת הרשמקול לפרקי זמן ממושכים.

עלינו להיות מודעים לעובדה שאנו צפויים למספר רב של סירובים ודחיות. בסקרי דעת קהל רגילים אנו צפויים לסירובים בשיעור של כ־20% או 25%, וכ־5% עד 20% אינם מסרבים מפורשות, אך או שאינם דוברים את השפה או שאינם מבינים מה רוצים מהם, וכן הלאה. כך שאחוז הנושרים עקב בעיות כאלה יכול להגיע ל־30% או 50%. בנוסף לכך, ייתכן שבפרוייקט הנוכחי מספר אנשים יעזבו באמצע הדרך, או שיכבו את מכשיר ההקלטה.

עלינו להיות ערים לכך שהאינפורמנטים, אלה שמקליטים, יסננו את הסיטואציות שהם מקליטים. כך, רבים מהם יכבו את הרשמקול כאשר יתחילו לריב עם בן/בת זוגם (אם יזכרו לעשות זאת). כמו כן, רוב האנשים יסגרו את מכשיר הרשמקול לפני שייגשו לשירותים, לחדר המיטות, לבנק או למשרד עורך הדין. כך אנו עלולים להחמיץ (או לקבל בתת־ייצוג) רבים ממצבי הדיבור, גם אם יהיו אנשים שיסכימו, עקרונית, להקליט הכול. ייתכן שיש לבקש מהמשטרה, או מכוחות הביטחון, לספק לנו כמה דגימות מהקלטות סודיות שלהם.

עלינו להיות מודעים גם לעובדה שיהיו מילים, מצבים, צורות דיבור, קהילות או תת־קהילות שלא ייכללו במידגם על־פי שיטות אלה. אם נרצה אותם בכל זאת (למשל, אם נרצה לוודא שיש ייצוג גם לדיבורם של עולי אתיופיה, או לעגת עבריינים) יהיה עלינו לדגום דגימת יתר את תת־האוכלוסיות הללו. הדבר לגיטימי, נוכל לעשות זאת, אך עלינו לתכנן זאת מראש.

לבסוף, עלינו להיות מודעים לכך שכל פרוייקט המערב הקלטה נרחבת כל כך, תובע הכנה רבה. הכוונה היא להכנה מינהלית ולוגיסטית מצד אחד, ולמציאת פתרונות לבעיות אתיות מסוימות לשביעות רצון צוות המחקר וגופי בקרה (ועדות הלסינקי וגופים דומים) מצד שני, שכן הקלטה, מטבעה, מעוררת סוגיות אתיות ומשפטיות רבות.