מבוא לבלשנות המאגר:
על חקר הקורפוסים בעולם ובישראל

עינת גונן

1. התפתחות בלשנות המאגר בעולם*

יותר מאלפיים שנה התמקד העיון הבלשני השיטתי בעולם המערבי בעיקר בשפה הכתובה הניתנת לחקירה פשוטה יחסית באמצעים טכנולוגיים שעד לאחרונה לא היה אפשר לחקור בהם שפות מדוברות. עם צמיחת הבלשנות המודרנית התהפכה הגישה, ובלשנים ידועי שם במחצית הראשונה של המאה העשרים הדגישו את החשיבות הרבה במחקר הלשון המדוברת (Chafe & Tannen 1987: 383). בעשורים האחרונים – בד בבד עם ההתקדמות הטכנולוגית ששכללה את מחקר הלשון המדוברת – אנחנו עדים לפריחה של חקירה בלשנית המבוססת על נתונים אותנטיים של דיבור (Leech 1991, 2000: 677, Johansson 1995: 243, Joseph 2008: 687). פריחה זו מתבטאת במספר המחקרים ובאיכותם, ובעקבותיה הולכות ומתגבשות נורמות מחקר חדשות: מחקר המושתת על עדויות מבוססות בשפה המדוברת, התפתחות תחומי חקירה חדשים ודיון ער באתגרים העומדים בפני חוקרי התחום החדש שהתגבש: בלשנות המאגר (corpus linguistics). תחום זה טומן בחובו הבטחות גדולות למדע הבלשנות, ובצדן אתגרים וקשיים.

היסודות לבלשנות המאגר המודרנית הונחו בסקר של שימושי האנגלית, שהתחיל בסוף שנות החמישים של המאה הקודמת באיסוף נתוני סקר אותנטיים בצורה לא אלקטרונית (Hardie & McEnery 2010: 384). בתקופה זו היה הקורפוס מרכזי להתפתחות התאוריה הבלשנית (Barlow 2011: 4): בשנים האלה תכנן רנדולף קירק את הסקר הגדול שלו על שימושי הלשון האנגלית (Survey of English Usage – SEU),[1] ובמקביל החלו פרנסיס וקוסרה לעבוד על המאגר המפורסם של

* מאמר זה נכתב בשנת 2011, ועודכן רק בפרטים אחדים. מאז נוספו מאגרים שונים בעולם, וגם המאגרים הקיימים שוכללו. סקירת המאגרים המובאת כאן היא בבחינת טעימה קטנה משלל החומרים המזומנים לחוקרי שפות זרות.

1 http://www.ucl.ac.uk/english-usage/about/history.htm.

בראון, שהטקסטים בו הם טקסטים כתובים משנת 1961.[2] ואולם כעבור זמן קצר היו הבלשנים האלה למיעוט בגל הבלשנות הגנרטיבית ששטף את התחום, גל שמייחסים לו השפעה משמעותית על העיכוב בהתפתחות חקר המאגר (ראו Barlow, שם).

האסכולה הגנרטיבית הסיטה את מוקד העיסוק הבלשני מגישה מונחית טקסט, שכן היא העמידה כהנחת יסוד את הרעיון שכל בלשן שהוא דובר ילידי, יכול להעלות דוגמאות משלו לתופעת לשון מסוימת ולקבוע אם מבנה נתון אכן קיים או אפשרי בלשון או שהוא איננו שייך למערכת הלשון של שפתו. פילמור במאמרו "'Corpus Linguistics' or 'computer-aided Armchair Linguistics'" רואה את עצמו בלשן שכזה, ומתאר בלגלוג עצמי קריקטורה של בלשן כורסה היפותטי: הוא יושב בכורסת המנהלים הרכה שלו כשעיניו עצומות וידיו משולבות מאחורי ראשו. מדי פעם הוא פוקח את עיניו, מזדקף, וקריאה מתמלטת מפיו: "וואו, איזו עובדת לשון מגניבה", חוטף את עפרונו, ורושם בקדחתנות. ואז הוא מפזז בחדווה במשך שעות סביב הממצא שחשף, אחוז התרגשות על הגילוי שמקרב אותו לידיעה מה היא הלשון באמת (Fillmore 1992: 35).[3] ברי שמדובר בקריקטורה, אבל בדומה לתיאורים סטראוטיפיים רבים, אפשר לראות בה גם גרעין של אמת. אל המתח בין הבלשנות החומסקיאנית לבלשנות המאגר נשוב בהמשך.

ובינתיים, בחלוף הזמן, ובד בבד עם ההתפתחות הטכנולוגית העצומה שהמחשב זימן לעולם המחקר, חלחלה ההכרה שמאגרי לשון עשויים לתרום תרומה מכרעת למחקר הבלשני, ויש לשקוד על פיתוחם. לפיכך החלו להיווסד מאגרים שונים, חלקם מוגבלים בהיקפם, אחרים עצומי היקף. מחשבה רבה מושקעת בתכנון מאגרי לשון, והמודל שעל פיו הם ערוכים מוקפד.[4] רוב המאגרים, בעיקר הגדולים שבהם, התמקדו בלשון הכתובה, אבל יש גם מאגרים העוסקים גם בלשון המדוברת או מיוחדים לה בלבד. הטעם לכך ברור: טקסטים כתובים נוחים יחסית לקליטה ולעיבוד אוטומטי במאגר (ראו למשל הדיון במאגר BNC אצל Biber, Conrad & Reppen 1998: 13 או Čermák 2009: 114), ואילו בעיבוד טקסטים דבורים עולות שאלות סבוכות כגון שימור ההקלטה, סוגיות של אתיקה, תעתיק ותיוג דקדוקי (Baker 2010: 13). בעזרת הטכנולוגיה המתקדמת מתרחב היקפם של מאגרי הלשון, ובעוד המאגרים הוותיקים כוללים מיליון מילים, מאגרים חדשים חותרים להיקף גדול בהרבה.

2 http://khnt.hit.uib.no/icame/manuals/brown/INDEX.HTM.

3 פילמור אינו חוסך שבטו גם מבלשני המאגר, וקריקטורה של בלשן שכזה, המשועבד רובו ככולו לפלטי המחשב, מוצגת גם היא.

4 ראו למשל Quirk & Svartvik 1979, Johansson 1995: 246–250, Čermák 2009, יזרעאל תשס"ב, יזרעאל, הרי ורהב תשס"ב, יזרעאל תשס"ג–תשס"ד ועוד חומרים באתר מעמ"ד.

ללשון האנגלית מיוחדים עשרות מאגרים מסוגים שונים. בשפות אחרות נראה שהתחום מפותח פחות (Parodi 2010: 69), אם כי מאגרים מכוּנָנים עתה גם בשפות אחרות, כגון ספרדית (Parodi, שם), צרפתית (Sankoff & Sankoff 1973, בן־טולילה תשמ"ט),[5] הולנדית (http://lotos.library.uu.nl/publish/articles/000113/bookpart.pdf), מאגר C-Oral-Rom לאיטלקית, ספרדית, פורטוגזית וצרפתית (Crestie & Moneglia 2005), צ'כית (http://ucnk.ff.cuni.cz/english/index.php), ערבית (Al-Sulaiti & Atwell 2006), רוסית (Sharoff 2006) ובשפות נוספות (Wilson, Rayson & Archer 2006). בנספח מובאת רשימה של כמה מאגרים חשובים לשפה האנגלית.

על יסוד המאגרים האלה נכתבות עבודות מחקר בנושאים בלשניים מוגדרים, מילונים או ספרי דקדוק מקיפים ורחבי יריעה, ובהם ספרו של סינקליר על יסוד מאגר COBUILD,[6] דקדוק לונגמן של בייבר ועמיתים,[7] ספרם הוותיק של קירק, גרינבאום וליץ' שהתבסס על הסקר לשימוש באנגלית,[8] תיאור מערכת הפועל של פרנסיס ועמיתים, שאמנם אזל בהוצאה לאור, אבל נמצא במלואו באינטרנט[9] ועוד.

העבודה על מאגרי הנתונים הגדולים הצמיחה למעשה ענף מחקר חדש: בלשנות המאגר. ענף זה מושתת על חקר הקורפוסים (corpora, בצורתו הלועזית), וביסודו עומדת החקירה של אוסף טקסטים כתובים או דבורים המתגבשים לכלל מאגר. לכאורה הגדרתו פשוטה יחסית, ואולם בחינת הגדרות אחדות שפרסמו חוקרים שונים מצביעה על מורכבותה.

Crystal (1997: 414) מגדיר קורפוס כדגימה מייצגת של הלשון שנאספה למטרות ניתוח בלשני. גם טוניני־בונלי (Tognini-Bonelli 2001: 2) מכוונת בהגדרתה לייצוג, וקובעת שקורפוס מוגדר כאוסף של טקסטים שיש הנחה שהם מייצגים לשון נתונה, והם מקובצים באופן היכול לשמש לניתוח בלשני (וראו עוד דיון מקיף בהגדרה שם: 55–52). בייבר ועמיתים (Biber, Conrad & Reppen 1998: 12) מגדירים קורפוס כאוסף **רחב** ו**מובנה** של טקסטים אותנטיים. עד כמה רחב הקורפוס – השאלה נותרת פתוחה. בארתס (Barthes 1967: 96) מציג הדגשים אחרים, ומגדיר קורפוס כאוסף **מוגדר** ו**תחום** של חומרים, הנקבע מראש על ידי המנתח באופן שיש בו מידה

5 וגם: http://www.uclouvain.be/valibel-corpus.html. ברשימות תפוצה שונות כגון corpora-List אפשר להתעדכן בחידושי המחקר בתחום.

6 Sinclair J. (ed.). 1988. *Collins COBUILD English Grammar*. New York: HarperCollins.

7 Biber D., Johansson S., Leech G., Conrad S. & Finegan E. 1999. *Longman Grammar of Spoken and Written English*. London: Longman.

8 Quirk R. *et al*. 1972. *A Grammar of Contemporary English*. London: Longman.

9 Francis G., Hunston S. & Manning E. 1996. *Grammar Patterns 1: Verbs*. London: HarperCollins.

מסוימת של שרירותיות בלתי נמנעת. אליו מצטרפים גם באואר וארטס (תשע"א: 32), הרואים בשרירותיות של הבחירה המופיעה בהגדרה של בארתס עניין הכרחי ברמת העיקרון. ואולם הקביעה שמאגר צריך להיות סגור ותחום אינה מקובלת על כל החוקרים, ונוכל למצוא מאגרים סגורים (סטטיים), או פתוחים (דינמיים). מאגר COBUILD, למשל, מייסודו של סינקליר הוא מאגר פתוח, טקסטים מוספים אליו כל הזמן, והוא בתהליך מתמיד של התרחבות (McEnery & Wilson 1996: 30). בכך נשלל גם עיקרון אחר שקבע בארתס (שם: 97–98), שסבר שיש לחתור להומוגניות בזמן המיוצג במאגר בדרך שתשלול היבטים דיאכרוניים בניתוח, שכן מאגרים המלקטים טקסטים מתקופות לשון שונות, עשויים לייצג רבדים מתפתחים של הלשון ולא רק את אבני הלשון המונחות זו בצד זו. חוקרים אחרים מזכירים מגוון סוגות, משלבים והקשרים (באואר וארטס תשע"א: 38). בהמשך לכך מעדיפה טוניני־בונלי (Tognini-Bonelli 2001: 6) לדבר על קורפורה ולא על קורפוס, שכן לדעתה אי־אפשר להעמיד קורפוס אחד שמתאים לכל המטרות, וממילא מכונֵן המאגר מתאים אותו לצרכיו. חוקרים אחדים מציינים את המדיום שבו הנתונים מקובצים כתנאי חשוב במאגר הבלשני המודרני, למשל Kennedy (1998: 3; הגדרות נוספות אצל באואר וארטס תשע"א: 31 ואילך). גם מהותה של בלשנות המאגר שנויה במחלוקת: האם היא בחזקת מתודולוגיה שיכולה להיות מסונפת לכל גישה בלשנית תאורטית (Thompson & Huntson 2006: 8, וראו גם Hardie & McEnery 2010: 385), או הרבה יותר מזה: ענף בלשני עצמאי הנשען על גישה פילוסופית חדשה לחקר הבלשנות ונושא עמו אופק מחקרי חדש? (Tognini-Bonelli 2001: 1, וראו גם Baker 2010: 6)

מקובל בקרב בלשני המאגר לראות בטקסטים המרכיבים את המאגר דוגמה לפארול (parole), בעוד התבניות העולות מהמאגר כולו מכוונות לתובנות באשר ללאנג (langue; Tognini-Bonelli 2001: 3). ואולם מידת הייצוג שתיוחס למאגר שנויה במחלוקת. מטבע הדברים השאיפה היא למידת ייצוג רבה ככל האפשר, ואולם הקהילה הבלשנית דחתה ב־1991 מהלך שנועד לקבוע כי מאגרים לשוניים ייחשבו כמייצגים את השימוש בשפה (Johansson 1995: 246). עם זה, יש המזהירים מפני מהלך מרחיק לכיוון ההפוך שבו משיקולי נוחות המאגר לא יהיה מייצג כלל וכלל. ג'והנסון (שם) מוסיף שהמאגר חייב להיות מייצג באופן כלשהו, בכך שישקף מגוון של סוגי טקסטים ובחירות לשוניות בשפה.

מידת הייצוג שאפשר לייחס למאגר קשורה לאופן תכנונו ולהיקפו. Cook (1995: 35) כותב שהכמות הנחקרת עשויה להשביח את האיכות, והיא מאפשרת ניתוח שמבחין בין צורות סדירות לבין צורות אידיוסינקרטיות. סינקליר (תשס"ב[1]: 7–8) עומד על ההבדלים בין טקסט לבין מאגר במתודולוגיה, בכלי הניתוח ובגישה

אל מאגר הלשון. הוא אמנם טוען שעקרונית אפשר להתייחס אל טקסט יחיד גדול מאוד כאל מאגר ולטפל בו בטכניקות מאגר ולא בטכניקות טקסטואליות, ולכן גודלו של המאגר אינו רלוונטי לאופי הטיפול, אבל הוא שב וקובע שהיקף קטן הוא מגבלה. אחרים סוברים שגודל המאגר הוא שיקול רלוונטי פחות, ואילו מידת הייצוגיות שבו מחייבת תשומת לב רבה יותר (באואר וארטס תשע"א: 37). נראה ששיקולים פרגמטיים מאלצים חוקרים רבים להתפשר בחקירתם, אף על פי שלמרבית הצרכים הבלשניים הם יזדקקו למאגרים גדולים מאוד. לפיכך מתחזקת ההנחה שבהיעדר מאגר גדול, גם מאגרים קטנים עשויים להספיק כדי לתאר רבים ממאפייני הדקדוק השכיחים (McCarthy & Carter 2001, וראו גם Ghadessy, Henry & Roseberry 2001). עם זה ברי לכול שככל שהמאגר יהיה גדול יותר, הוא יהיה בעל אופי מייצג יותר ובעל פוטנציאל גבוה יותר לשקף מצע לתיאור חוקיות לשונית.

מרבית המאגרים בנויים על פי כמה פרמטרים כגון ערוץ (דבור או כתוב, דבר מה שנכתב לצורך קריאתו), תחום (אמנותי, ביתי, דתי, חינוכי וכיו"ב), תפקיד (לשכנע, להביע, ליידע וכיו"ב; באואר וארטס תשע"א: 32–33).

מאגר יכול להיות מתויג (annotated) או בלתי מתויג (McEnery & Wilson 1996: 32). התיוג מאפשר שליפה וניתוח על פי הקטגוריות שנבחרו, למשל גופים, שורשים, תחיליות, סופיות, חלקי דיבור ומגוון קטגוריות תחביריות, ומקל מאוד את העבודה הבלשנית.

הפוטנציאל הגלום במאגרי לשון הוא עצום, ונוגע לתחומים רבים, לא בהכרח בלשניים בלבד (ורום תשס"ב). בהיבט הבלשני ברי שמתוך הטקסטים עולה תמונה אותנטית של הלשון, ומתאפשרים מחקרים כמותיים העשויים לצייר תמונה חשובה של שכיחות צורות ומבנים בלשון. יתרה מזאת, באמצעות חקירת קורפוס מוצג מכלול של שונות לשונית, בלי לסמוך רק על מידת בקיאותו של הבלשן בווריאציות השונות (באואר וארטס תשע"א: 34; De Monnink 2000: 133–134, Čermák 2009: 113). גם בתחום ההוראה, הוראת שפת אם והוראת שפות זרות, מקובל מאוד להיעזר היום בקורפוסים, ויש הרואים בהם כלי לשינוי פני ההוראה (Tognini-Bonelli 2001: 14, McEnery & Xiao 2011, ופרסומים רבים אחרים).

2. התבססות התחום

בלשנות המאגר משמשת קרקע פורייה לחקירה כמותית ואיכותנית בענפי המחקר השונים, ופריחה של ממש בחקר הלשון ניכרת בכל התחומים בשתי שיטות המחקר הללו. בעיקר נחקרים בעולם היבטים של תחביר, של שיח ושל אינטונציה בשיח הדבור, בעוד תחומי דקדוק אחרים נחקרים פחות (Aarts 2002: 10). בארלו מציין

שלושה תחומים שבהם השפעתה של בלשנות המאגר בולטת במיוחד: חקר הקולוקציה, העיסוק בשכיחות ובצורות טיפוסיות של מבעים (במקום צורות אפשריות של מבעים) ומחקרים כמותיים על שונות לשונית (Barlow 2011: 7–8). מטבע הדברים יהיו תחומי מחקר שייטו לכיוון הכמותי (למשל פונטיקה, פונולוגיה ומורפולוגיה), ואילו אחרים ייטו לכיוון האיכותני (תחומי השיח, הסמנטיקה וחלקים מן התחביר – וגם בהם יוצגו לעתים מודלים סטטיסטיים מורכבים). יש בלשנים המדברים על גישה משולבת (multi-method approaches) לניתוח הקורפוסים (Schmied 1993, McEnery & Wilson 1996: 63), אך לא כאן המקום להרחיב.

ככל שמתרחב העיסוק בתחום, צומחים להם ענפי משנה, למשל אסכולות של בלשנות מבוססת מאגר (corpus based linguistics), בלשנות מונעת מאגר (corpus driven linguistics), בלשנות חישובית ועוד.

2.1 בלשנות מונעת מאגר ובלשנות מבוססת מאגר

טוניני־בונלי מסבירה את הבדלי הגישות בין בלשנות מונעת מאגר למבוססת מאגר: גישה מבוססת מאגר נוגעת לסוג מתודולוגיה שאין בו מחויבות שיטתית לנתונים כמכלול. הקורפוס משמש בדרך כלל כדי לתת תוקף לקטגוריות שהונחו מראש או כדי להשלים את התאוריה בממד ההסתברות. הגישה מונעת המאגר, לעומת זאת, משמשת לבחינה הוליסטית של הנתונים כמכלול מתוך כוונה לתארם תיאור מקיף על יסוד העדויות. הקורפוס אינו משמש בה למתן תוקף להנחות מוקדמות. בעזרת בחינת העדויות, לרבות בחינת מידת השכיחות של הדגמים או היעדרם של דגמים אחרים, אפשר להעמיד קטגוריות בלשניות חדשות (Tognini-Bonelli 2001: 81, 84, 87, יזרעאל תשס"ה: 336).

לפי טוניני־בונלי (שם: 65–100) וסינקליר (תשס"ב[1]: 11), הבלשנות מונעת המאגר אינה משתמשת בטקסט מתויג אלא מעבדת ישירות את הטקסט הגולמי כך שדפוסים הניכרים מתוך הטקסט הגולמי צצים ועולים לעיני החוקר. מאחר שתיוג מראש מגביל את שאלות המחקר הנשאלות ומצמצם את רוחב היריעה של התיאור הלשוני, מוטב לדעתם לערוך את עיבוד הלשון מתוך התבוננות בטקסט. לבלשנות מונעת המאגר נדרשים מאגרי לשון עצומים בגודלם, משום שהיא דורשת היקרויות רבות ככל האפשר של היחידות שהיא מטפלת בהן (סינקליר תשס"ב[1]: 13).

מכיוון שהבלשנות מונעת המאגר היא תחום חדש יחסית שעדיין מבקש את דרכו, ההבחנות הנחרצות עדיין לא נתקבעו בתחומים רבים, ואפילו מחקרים מונעי מאגר על פי רוב אינם נאמנים לגמרי לעיקרון המתודולוגי שבבסיס השיטה. כך

למשל מחקרים רבים מתחילים בתחושה שיש לחוקר בנוגע למילה מסוימת, והמחקר עשוי להישען על תאוריה מבוססת (Gries 2010: 328–329, Barlow 2011: 5).

ההבחנות שטבעה טוניני־בונלי בשנת 2001 התפשטו במהירות ועוררו הדים רבים. במידה מסוימת דומה שנוצרו שני מחנות שונים: בצד אלו שמיהרו לאמץ את עקרונות הגישה מונעת המאגר, היו חוקרים אחרים שהסתייגו מחידוד ההבדלים. Gries (2010: 338) במאמרו "בלשנות המאגר והבלשנות התאורטית: יחסי אהבה – שנאה? לא בהכרח" הציע לחשוב מחדש על הניגוד שבין בלשנות מבוססת מאגר לעומת בלשנות מונעת מאגר, וקרא לצמצמו, וודאי שלא לחשוב עליו במונחים של קרב בין "אנחנו" כנגד "הם". גם Hardie & McEnery (2010: 385, 389–390) סוברים שההבחנה בין שני המחנות אינה שימושית כלל וכלל, ולכן הם תומכים בשיתוף פעולה רעיוני ומתודולוגי בין שתי האסכולות. במאמרם הם מנסים לטשטש את היריבות בין שני המחנות, ומצביעים על שיתוף פעולה הדוק בתחומים אחדים.

2.2 הבלשנות החישובית

בצד בלשנות המאגר מתבססת לה הבלשנות החישובית (computational linguistics), שהיא תחום מחקר המקשר בין בלשנות ובין מדעי המחשב. שולי וינטנר (תשס"ב: 35) מסביר את העיסוק בתחום:

> מזווית ראייה אחת, זהו תחום המיישם שיטות ותוצאות של מדעי המחשב בבלשנות, על מנת לחקור שאלות יסוד של הבלשנות כגון מה אנו יודעים כאשר אנו "יודעים" שפה כלשהי, איך ומתי אנו משתמשים בידע הזה, כיצד אנו רוכשים אותו וכו'. מנקודת הראות השנייה, בלשנות חישובית מיישמת ידע ושיטות של הבלשנות במדעי המחשב, על מנת לייצר תכניות מחשב המבינות דיבור אנושי, מתרגמות משפה טבעית אחת לאחרת, ובאופן כללי מתקשרות באופן מילולי עם בני אנוש בדרכים המותאמות לאנשים ולא למחשבים. לעתים משתמשים במונח **עיבוד שפות טבעיות** (natural language processing) ביחס לנקודת המבט השנייה.

הזיקה בין הבלשנות החישובית לבלשנות המאגר ברורה. אבל מעבר לזיקה המדעית בין שני התחומים עומד גם עניין מעשי־כלכלי: עם התקדמות הטכנולוגיה מאגרים רחבים של דיבור מוקלט נדרשים לצורך הכנת תכנות המפענחות דיבור טבעי, ומצאי רחב של אלופונים מדיאלקטים שונים צריך להיבחן כדי להגביר את שיעורי הזיהוי של דיבור אוטומטי (Pineda *et al.* 2010: 349). לפיכך מעורבות לפעמים גם חברות עסקיות בפיתוח מאגרי לשון.

דומה שהעיסוק בפענוח ממוחשב של השפה, המושתת בעיקרו על סטטיסטיקה ולכן מצריך היקף עצום של קורפוסים דבורים, מטריד כמה מבלשני המאגר, והם יוצאים להתקפה כנגד "מהנדסי הלשון", המכנים עצמם "בלשנים". סינקליר (תשס"ב[1]: 3–4) סובר שההישגים בנושא הקשר שבין הלשון והמחשוב מאכזבים, ואינם מבטיחים רבות, ומסויג מן האפשרות שקוד הלשון יפוצח בסופו של דבר, ותרגום אוטומטי של לשון חופשית ייעשה בהצלחה מלאה (וראו גם Sinclair 2004: 185ואילך). ניתוח אופטימי קצת יותר נמצא אצל אנשי הבלשנות החישובית, למשל אצל וינטנר (תשס"ב: 38), אם כי גם הוא סובר שאין בידינו כרגע הכלים להבנת שפות טבעיות והחלת יישומים מורכבים. על המחלוקת העזה הניטשת בין שני ענפי הבלשנות האלה אפשר ללמוד גם מעיון אצל Wilks 2010 ו־Teubert 2010.

2.3 עד כמה מייצגים ממצאי המאגרים?

אחד הנושאים העיקריים המעסיקים את חוקרי התחום הוא תקפותם של הממצאים, נושא הניצב ביסוד התחום כולו. הדיון הזה מחזיר אותנו למתח בין הבלשנות החומסקיאנית לבלשנות המאגר, ונרחיב כאן בכמה עניינים. כאמור, התנגדו חומסקי והבלשנים התאורטיים התנגדות של ממש לבלשנות אמפירית (Meyer 2009: 210), עד כדי קביעה נחרצת של חומסקי שאין דבר כזה כבלשנות המאגר (באואר וארטס תשע"א: 34). הוויכוח בין שני הזרמים אינו ויכוח של יוקרה בלבד או של גישות תאורטיות, שכן הוא נוגע בשאלה הניצבת בלב לִבָּן של שתי האסכולות הבלשניות האלה: מה יכול לסרטט תמונה נאמנה יותר של הלשון: קורפוס תחום, יהא גדול ככל שיהא, או מערכת הלשון החובקת העומדת לרשות הבלשן הילידי כדובר השפה, ותאורטית היא כמעט אינה מוגבלת בהיקפה.

חומסקי מייצג גישה הטוענת שדווקא דגימה של הלשון עלולה להטעות וליצור מצג שווא על דמותה, שכן הקורפוס מוגבל ואינו יכול להבטיח תיאור מלא ושלם של תופעה נחקרת. ובעיקר הדבר תקף למבנה שכיח שעשוי להופיע מעט או להישמט לחלוטין, לעומת מבנה נדיר, שעשוי להזדמן בקורפוס נתון דווקא בשכיחות גבוהה (Greenbaum 1984: 193, McEnery & Wilson 1996: 61).

מהצד האחר של המתרס עומדים בלשנים סטרוקטורליסטיים, הסוברים שעל הבלשנות להישען על נתונים אמתיים ולא מומצאים (ראו למשל De Monnink 2000: 133). היטיבה להגדיר את מטרות התיאור הלשוני Ochs (1979: 43): יש לתאר את מה **שיש** בניגוד למה **שאמור להיות**. במאמרו הידוע של סינקליר "Trust the text", הפותח קובץ מאמרים שחיבר הנושא שם זה, ומבוסס על הרצאה שנשא

בשנת 1990, סינקליר קורא לחוקרים להיות קשובים למה שהטקסט מספר ולבטוח בו (Sinclair 2004: 23).

חשוב להעיר כאן שהטענה על מידת תקפותם של ממצאי המחקר והשאלה על מידת הייצוג שבהם תקפה בכל שיטת מחקר המבוססת על ממצאי לשון. מסד נתונים של שפה אותנטית הוא רק דרך חקירה אחת המזומנת לחוקרי הלשון, ושיטות מחקר אחרות המבוססות על מבדקים מתוכננים רווחות גם הן. גם בדרכי המחקר האלטרנטיביות עולות שאלות הנוגעות למידת הייצוג שיש לייחס לממצאים, והדבר מותנה כמובן בשאלת המחקר, במספר האינפורמנטים שנבחרו ובדרך בחירתם, במספר הפריטים שנבדקו וגם בממצאים עצמם (ראו גונן תשע"ג).

בלשני המאגר אינם מבטלים כלל וכלל את הטענה שגם מאגר גדול של נתונים בן כמה מיליוני מילים לפחות אינו מייצג בהכרח את הלשון הנוהגת, ובעיקר לא את שכיחות הצורות שבה, ורבים מהם עסוקים במחקריהם במציאת פתרון לבעיה. באופן פרדוקסלי מעט הם מונחים דווקא על ידי אותו גורם שעליו העמיד חומסקי את תורתו: חושם הלשוני כדוברים. לכן הם נדרשים לבחינה מתמדת של הנתונים ולבדיקה אם אין הממצאים שעלו סותרים בבוטות את מציאות הלשון. פילמור מנסח זאת בבהירות: היכולת לשפוט שקורפוס אינו גדול דיו לייצג תופעת לשון נבדקת היא יכולת המבוססת על ההכרה שדבר מסוים שהבלשן יודע אינטואיטיבית על הלשון כדובר ילידי, אינו מוצג בקורפוס (Fillmore 1992: 38).

לכאורה עולה כאן סתירה מניה וביה שהרי אם נבטח בטקסט, כמאמרו של סינקליר, ועוד יותר מזה, אם נניח לקורפוס עצמו להוביל אותנו בחקירתנו בבואנו אליו חפים מעמדות מגובשות ומתאוריות שנרצה לבחון, כפי שמציעה טוניני־בונלי, חלוצת הגישה הבלשנית מונעת המאגר – כיצד נוכל לשלב את האינטואיציה שלנו עם עובדות של ממש העולות מן הטקסט?

ואכן גם טוניני־בונלי מתלבטת: מצד אחד היא אכן מעלה ספקות על תפקיד ההנחות האינטואיטיביות בחקר הלשון וסוברת שהן אינן מייצגות נאמנה כלל וכלל את שימושי הלשון באמת; מצד אחר היא קובעת שהאינטואיציה היא עדיין גורם חיוני שיהיה בעל משקל גדול למשל בבחירת תופעה שבלשן מתאר, ובסופו של דבר היא תמלא תפקיד חשוב כשנתוני הקורפוס ייבחנו ויוערכו (Tognini-Bonelli 2001, וראו דיונו של Aarts 2002: 9). מכל מקום, דה־מונינק קובעת שאף על פי שהאפשרות לבסס מחקר כמותי על קורפוסים היא אחת מנקודות החוזקה של המאגר, יש לנהוג בנתונים זהירות מופלגת בבואנו לפרשם (De Monnink 2000: 138). לפיכך היא מציעה במחקרה על הצירוף השמני באנגלית לשלב בין שיטות חקירה. לטענתה, כאשר הניסוי מתוכנן בקפידה, נתונים הנשאבים מניסויים מתוכננים (elicitation data) עשויים להיות מקור משלים חשוב לנתונים המתקבלים ממאגר.

את תקפות הממצאים בזיקה למתח שבין הידע האינטואיטיבי של הבלשן לבין התמונה העולה מקורפוסים בחנו ברג ועמיתים (Bergh, Seppänen & Trotta 1998). הם בדקו את ההבדלים שבין חיפוש בקורפוסים גדולים ייעודיים של הלשון האנגלית לבין חיפושים במאגרים עצומים של האינטרנט – וזאת בזיקה לתחושה האינטואיטיבית של הדובר. מסקנות המחקר שלהם (שם: 52–53) מחדדות שתי נקודות:

א. אם צירוף לשוני נקרה בקורפוס באופן חוזר, אפשר להסיק שהוא חלק מן הלשון. אם איננו מזדמן, הבלשנים יכולים לסמוך רק על התחושה הפנימית שלהם אם לפנות לאימות אמפירי באמצעות שימוש בקורפוס גדול יותר או במבדקי אינטואיציה.[10]

ב. חשוב שהבלשן יהיה מודע לפוטנציאל הגלום במאגרים תחומים או פתוחים ולהשתמעות הנגזרת מגודלם; בלשון הכתובה האינטרנט יכול לשמש מאגר משלים למאגרי הלשון הקיימים.

בעקבות העיסוק האינטנסיבי בתחום אפשר להבחין בשינויי תפיסה גם בקרב חלק מהמתנגדים להצבת הטקסט כמקור החקירה. פילמור (Fillmore 1992: 35), שכאמור רואה בעצמו "בלשן כורסה", מדווח על מסקנותיו מהתנסות עם עבודה במאגרים, ומסיק שתי מסקנות: האחת היא שלדעתו לעולם לא יימצא מאגר לשון שיכיל מידע מקיף דיו על כל צורות הלשון האנגליות שהוא מעוניין לחקור; האחרת היא שכל קורפוס שהזדמן לו לבדוק, ואפילו קטן, לימד אותו עובדות לשון שלא יכול היה לדמיין את קיומן בכל דרך אחרת. לפיכך הוא מסיק ששתי הגישות משלימות זו את זו (שם).

ואכן באמצעות חקירת קורפוס מוצג מכלול של שונות לשונית מבלי לסמוך רק על מידת בקיאותו של הבלשן בווריאציות השונות (באואר וארטס תשע"א: 34, De Monnink 2000: 133–134).

4. בלשנות המאגר העברית

4.1 חקר העברית המדוברת: תקופת הבראשית

בעוד בעולם הרחב הנצו בשנות החמישים והשישים זרמים בלשניים אחדים שהעמידו עקרונות מדעיים ברורים בתשתית החקירה, וגם בלשנות המאגר צעדה את צעדיה הראשונים, הייתה הבלשנות העברית בת ימינו הרחק מאחור. אמנם בחקר העברית הקלאסית המשיכו חוקרי העברית לפתח את מסורת המחקר הענפה ואף לכונן מאגרים מתקדמים ללשון הקדומה, ובראשם המילון ההיסטורי ללשון העברית

10 ואחרים יציעו מודלים מעוררי מחלוקת להתמודד עם המצב הזה (ראו ביקורתם של Gale & Church 1994).

שהושתת מראשית דרכו על אמצעים ממוכנים, ואולם באשר למחקר סינכרוני של הלשון הכתובה והמדוברת נחלקו הדעות.

בתקופות הראשונות של תולדות הלשון העברית המדוברת בארץ ישראל לא תוארה הלשון העברית המִתְחַיָּה תיאורים בלשניים לפי אמות המידה הבלשניות המדעיות שנהגו בקהילה המדעית באירופה (ראו Fishman 1974: 9–11). במחקר העברית ניכר עירוב תחומים בין הבלשנות לקביעת נורמה (Weinberg 1966: 40–41, Rosén 1977: 40–42), והחוקרים והמורים התמקדו בדקדוק העברית החדשה מתוך גישה נורמטיבית בעיקרה, והסתייגו מתיאור אובייקטיבי שלה (שורצולד תשס"ב: 57, כאן תשס"ב: 279). בשל האופי האידאולוגי של מעשה תחיית הלשון והמציאות המיוחדת שאפיינה את ימי טרום המדינה ואת שנותיה הראשונות, היו הבלשנים חדורי אידאולוגיה ציונית־לשונית, דבר שהשפיע על שיקול דעתם בנוגע לדרכי החקירה הלשונית. בתיאור הלשון כמות שהיא הם ראו השפעה שלילית על עיצובה, ואפילו מתיאור שפתם של הילדים, המשופעת "שיבושים", הסתייגו הבלשנים הטהרנים (בר־אדון תשכ"ג: 22).

בשנות החמישים נקלעה קהיליית הבלשנים העברית לסערה גדולה, והמתח בין עיסוק בלשני מתאר לבין עשייה נורמטיבית חלוצית התגלע במלוא עוזו. הקונפליקט הזה כבר תואר בפירוט בכמה פרסומים (ראו להלן), אבל נשוב ונעמוד כאן על עניינים הנוגעים להשפעותיו על המחקר כיום. ב־1950 התפרסם מחקרו של Weiman (1950) על הרכיבים הזרים והמקוריים בעברית המדוברת. מחקר זה, שנערך בניו יורק בשנים 1946–1949 והוקלטו בו דוברים ילידיים, נחשב מחקר חלוצי בשל התיאור הסטרוקטורליסטי שבו (רוזן תשט"ו: 5–6), אך לרבים מן הבלשנים היו השגות על תיאור העובדות שבו ועל ערבוב בין היבטים סינכרוניים ודיאכרוניים.

שני אישים מרכזיים עמדו בראש המאבק למחקר מתאר של העברית המדוברת: חיים רוזן וחיים בלנק, והם פרסמו מחקרים בנושא וטקסטים מתועתקים (רוזן תשט"ז, בלנק תשי"ז, Blanc 1964; וראו גם דיון מפורט אצל Kuzar 2001: 152 ואילך, ורשף תשס"ד: 34 ואילך). רוזן אף פצח בפולמוס פומבי בבימות מחקר שונות עם בלשנים נורמטיביים, ובראשם זאב בן־חיים, מכונן מפעל המילון ההיסטורי של האקדמיה ללשון העברית (רוזן, שם ובפרסומים נוספים, בן־חיים תשי"ג). אנשי הנורמה מצדם חששו שתיאור עובדות הלשון הנוהגת יהיה צעד חשוב לקראת קבלתן כמקובלות ותקינות, כפי שקרה בעבר בשפות אחרות בעולם (רשף תשס"ד: 37).

בצד המודעות הבלשנית המתפתחת תרמו גם אסכולות המחקר המתפתחות באוניברסיטאות לפתיחת התחום למחקר סינכרוני, ובראשם החוגים ללשון העברית (גושן־גוטשטיין תש"ם: 191, הערה 13, בר־אשר תשס"ג: 230–234) ולבלשנות (Rosén 1977: 42–43) באוניברסיטה העברית. אך ההתקדמות הייתה אטית: בהדרגה

החלו להיכתב תיאורי לשון של יחידים על הלשון המדוברת, בעיקר בקרב בלשנים עבריים שלמדו לימודים אקדמיים מחוץ לגבולותיה של ישראל (ובהם חיים בלנק, אורה שורצולד, שמואל בולוצקי, משה חן, הולי זמילהוף־זלסקו), ומסורות מחקר סינכרוניות המתבססות על עובדות מתועדות ונטולות מטען של אידאולוגיה לשונית חלחלו לדפוסי החקירה. בצדם פעלו חוקרים אחרים שהחלו לעסוק בעברית המדוברת, ואולם לא על יסוד עדויות מוצקות אלא על פי תחושות אינטואיטיביות באשר לצורות הלשון הרווחות (שורצולד תש"ע: 324). בר־אשר מנסח את העניין בחריפות רבה: "קשה לי להתעלם מטענתם של רבים שכמה מן העוסקים בעברית החדשה על פי השיטה הדסקריפטיבית כבר יצא טבעם כלא־יודעי עברית, אם לנקוט לשון מתונה. הם בוראים ובודים לעתים משפטים שדובר לא שמע וכותב לא כתב" (בר־אשר תשס"ט–תש"ע: 11).

4.2 כינון מאגר לאומי: פרויקט מעמ"ד

על הרקע הזה צמחה בסוף שנות התשעים של המאה הקודמת יזמה חדשה של קבוצת חוקרים ובראשם שלמה יזרעאל מאוניברסיטת תל־אביב, בנימין הארי מאוניברסיטת אמורי שבארה"ב, מירה אריאל מאוניברסיטת תל־אביב, ג'ון דו בואה מאוניברסיטת סנטה ברברה שבארה"ב וגיורא רהב מאוניברסיטת תל־אביב, לכונן מאגר מקיף ללשון העברית המדוברת, מאגר פתוח לכול. המודל אשר על פיו תוכנן פרויקט מעמ"ד (מאגר העברית המדוברת בישראל) ראה לכלול מדגם מייצג של דוברי העברית בישראל, ובו הקלטות ספונטניות ותוספות משלימות בהיקף כולל של חמישה מיליון מילה (ראו אתר מעמ"ד). המאגר תוכנן בהקפדה יתרה, כך שיוכל לענות על צורכי מחקר שונים ולהיות מסד נתונים מייצג בדומה למאגרים מתקדמים אחרים בעולם (יזרעאל, הרי ורהב תשס"ב, יזרעאל תשס"ב, יזרעאל תשס"ג–תשס"ד ורשימה מפורטת באתר מעמ"ד, וראו גם Čermák 2009: 114). בעזרת מימון נדיב מאוניברסיטת תל־אביב ומימון נוסף מאוניברסיטת אמורי הוחל במחקר חלוץ, ובו הקליטו בשנים 2001–2002 מכוני מחקר המתמחים בסטטיסטיקה מקצועית שיחות הלקוחות מחיי היום־יום של מקליטיהן. להקלטות האלה התווספו טקסטים נוספים שהקליטו אנשי צוות מעמ"ד ותלמידיהם, ובסך הכול הועמדו כחמישים הקלטות בנות 8 עד 16 שעות כל אחת. ואולם למרות הצלחת מחקר החלוץ לא עלה בידו של צוות מעמ"ד לגייס את התקציב שהצריך פרויקט בסדר גודל כזה, ובעקבות ביטול הפרויקט לתמיכה במפעלים לאומיים במדעי הרוח מטעם האקדמיה הישראלית למדעים בשנת 2003, הוקפא המפעל.

מראשית הדרך ועד היום ניתנה לחוקרים ולסטודנטים גישה להקלטות שונות, וקטעים רבים תומללו בסמינרים לחקר העברית המדוברת באוניברסיטת תל־אביב. טקסטים אחרים תומללו או תועתקו לשם מחקרים רחבי היקף, בעיקר במסגרת עבודות לתואר שני ושלישי. באתר מעמ"ד מפורסמת רשימה מקיפה של מחקרים שהתבססו על מאגר העברית המדוברת בישראל. נוסף על כך בסיוע מענק מחקר של הקרן הלאומית למדע שהוענק לאסתר בורוכובסקי בר־אבא אף הוגשו לחוקרים בתחום כמה שעות מתומללות מתוך מחקר החלוץ מעובדות בתכנת ELAN, תכנה המאפשרת עבודה על תמליל צמוד שמע, ומשביחה לאין ערוך את כלי המחקר. החומר הזה הוא ביסוד עבודתם של חוקרי הסדנה שעליה מבוססים פרקי ספר זה [ראו פירוט במאמר "מסד הנתונים למיזם 'חוקרים עברית מדוברת'" מאת שלמה יזרעאל (יזרעאל תשע"ו)]. הטקסטים האלה וטקסטים נוספים מפורסמים באתר מעמ"ד במרשתת.

4.3 מאגר השיח של אוניברסיטת חיפה

בשנת 1994 החלה יעל משלר, חוקרת שיח מאוניברסיטת חיפה, לאסוף באופן שיטתי הקלטות ותמלילים, והם הועלו לאתר **קורפוס השיח העברי הדבור של אוניברסיטת חיפה**.[11] במאגר מקובצים חומרים שהוקלטו בשנים 1994–2014 (וכן שיחה שהוקלטה ב־1986), והוא מוסיף להתעדכן. האוסף מורכב מהקלטות אודיו של שיחות בין סטודנטים יהודים דוברי עברית ילידיים באוניברסיטאות ובמכללות שונות בארץ באינטראקציות בלתי פורמליות עם חבריהם ועם בני משפחותיהם. המשתתפים נפגשו בקבוצות בנות 2–5 דוברים עם חברים ועם בני משפחה שעמם נפגשו ממילא (כלומר לא באופן מבוים ושלא בנוכחות החוקרת), ונתבקשו להפעיל לכל אורך הפגישה מכשיר הקלטה במקום מרכזי בחדר (משלר תשס"ט: 101). בקורפוס מובאות 243 שיחות יום־יום מתומללות, שהן 11 שעות, 8 דקות ו־27 שניות של שיח בין 701 דוברים שונים, בד"כ 2–5 דוברים בשיחה. המאגר הוא מאגר דינמי, ומתוכננות להיכנס אליו שיחות נוספות. בשלב הזה הוא אינו פתוח לקהיליית החוקרים כולה, והוא עומד לרשות חוקרים עמיתים ותלמידים באישורה של משלר (רשימה של עבודות מחקר שנכתבו בהדרכתה מפורסמת באתרה).

ראוי לציין כאן את העבודה החשובה הנעשית בחקר לשון הילדים בניצוחן של רות ברמן, דורית רביד, אותי בת־אל וגלית אדם מאוניברסיטת תל־אביב ואחרים. תחום זה, שבדרך כלל אינו מסונף למחלקות ללשון העברית, מצמיח עבודות חשובות בחקר השפה המדוברת על יסוד הקלטות ומתוך מתודולוגיות מובנות שנתגבשו

11 קורפוס אוניברסיטת חיפה: http://weblx2.haifa.ac.il/~corpus/corpus_website.

בעולם האקדמי. בתחום זה רווח שיתוף הפעולה בין חוקרים במחקרים בישראל ובפרויקטים בין־לאומיים, והוא זוכה פעמים רבות גם לתקציב נדיב. טקסטים עבריים שונים – של ילדים ושל בוגרים – מזומנים במרשתת כבר עכשיו לרשות החוקרים גם בצורת השמע שלהם וגם בתעתיק CHILDES,[12] ומחקרם מניב מסקנות חשובות באשר ללשון המדוברת הנוהגת היום.

5. בלשנות המאגר העברית – אתגרים ומחשבות לעתיד

למרות ההתקדמות הגדולה שחלה בתחום בעיקר בשני העשורים האחרונים (שורצולד תשס"ח: 439–440), עדיין רחוקה העברית מהמשאבים העומדים לרשות חוקרי לשון בעולם. הבעיה היסודית של היעדר מאגר מוסכם, מתומלל וממוחשב מקשה מאוד על מחקר מבוסס עדויות של העברית המדוברת (יזרעאל תשס"ב: יג, בורוכובסקי תש"ע: 9), מעכבת אותו, ומנוטת אותו לתחומים שבהם אפשר לחקור גם בלא מסד נתונים רחב־היקף.

זוהי המכשלה העיקרית העומדת בדרכם של חוקרי העברית המדוברת, והיא ברורה וזועקת – אך אין היא היחידה. בעיה אחרת נוגעת להיעדר מסורת מחקר מבוססת. בבלשנות העולם טרם התגבשו מתודולוגיה ודרכי חקירה מוסכמות של מאגרים (Barlow 2011: 5), קל וחומר בבלשנות העברית. ואכן, בשעה שתחומים רבים של חקר הלשון העברית הקלאסית נהנו ממסורת מחקר מפותחת בת מאות שנים, היה מחקר הלשון המדוברת זנוח יחסית, ולא נתגבשה בו מסורת מתודולוגית אחידה לחקירה. העיסוק המקיף בעברית הקלאסית פעל בשני כיוונים מנוגדים: מצד אחד, הרקע הקלאסי חשוב מאוד להבנת רבים מתחומי הלשון גם בעברית בת ימינו, וידע נרחב שלו ושל תהליכי לשון שחלו ברבדים קדומים של הלשון העברית מעשיר ומעמיק גם את ההבנה על המתרחש בעברית בת ימינו. מהצד האחר, העיסוק הקלאסי של בלשני מופת ידועי שם בקורפוסים כתובים עתיקים בעלי היקף סגור ותחום, שלחלקם כבר נכתבו ספרי דקדוק ומילונים, יש השפעה משתקת על חקירת העברית הדבורה. לקורפוס עתיק יש התחלה ויש סוף, הוא מזמין מחקר איכותני וכמותי מבלי שיעלו כלל שאלות על מידת הייצוג שיש בחקירה הזאת באשר לשפה שנהגה בתקופה המתוארת, ומבלי שיפקפקו בחשיבות המחקר או בדרכי הצגתו בדמות רשימות רשימות. מכאן שלא זו בלבד שבלשני העברית החדשה בכלל והעברית המדוברת בפרט נדרשים לגבש מתודולוגיות המתאימות לצורכי מחקרם, עליהם גם להשתחרר מִצֶלֶם של דפוסי מחקר קודמים בעלי יוקרה ומוניטין, שאינם מתאימים כלל ועיקר לאופי החקירה של לשון מדוברת נוהגת.

12 http://childes.psy.cmu.edu.

האתגרים המתודולוגיים נוגעים גם להנחת תשתית חדשה של טרמינולוגיה ההולמת תיאור לשונות מדוברות, ושאלות יסוד כגון מהי מילה בשפה המדוברת, מהו משפט או מבע ושאלות רבות אחרות שטרם נוסחו, מצריכות עיון מחודש. בייחוד הדבר בולט בתחומי לשון בעלי מסורת מחקר ותיקה בלשון הכתובה, לעומת תחומים אחרים, חדשים יחסית במסורת העברית או כאלה המייחדים מלכתחילה בעיקר את הלשון המדוברת (כדוגמת הפרוזודיה או הפונטיקה), שבהם עול מסורת המחקר של הטקסט הכתוב מכביד פחות. גם כאן עומדת לפנינו סוגיה מורכבת, ומתחדדים הלבטים אם להשתמש במונחים מוכרים גם אם אינם מדויקים, או שמא לקבוע מונחים חדשים בעלי רמת דיוק רבה יותר, שאינם נושאים מטען של מחקרי העבר (ראו למשל לבטיו של Chafe 1987: 21–22, ודיונה של ורטהיימר תשס"ד). ואולם לא רק במינוח עסקינן אלא גם בתפיסות מהפכניות לחקר הלשון הדבורה. סינקליר קרא למהפכה כוללת כבר לפני 25 שנה ויותר (Sinclair 1985: 252, וגם סינקליר תשס"ב[2]: 32), ומאז חלו שינויים רבים – אך מהפכה של ממש בתפיסות הנוגעות לחקר הלשון העברית, כמו גם לחקר לשונות מדוברות אחרות, טרם התחוללה.

אפשר להרהר מדוע מהפכה שכזאת עדיין לא התרחשה למרות האמצעים המפותחים בהרבה העומדים לרשות חוקרי לשונות אחרות ומסורת המחקר המפותחת יותר המסורה להם. נדמה לי שמעבר לסגירות המחשבתית ולציפייה למנהיג מהפכן שינהיג שינוי רדיקלי (ומכאן גם להתלבטות אם מהפכות רדיקליות, דרכן להצליח לאורך ימים), עולה ביתר שאת השאלה על מהותה של הלשון המדוברת וזיקתה ללשון הכתב: האם יש לה קיום עצמאי בצד הלשון הכתובה, כזה התובע עולם מושגים חדש, או שמא היא אופנוּת בלבד של לשון תשתית אחת מנטלית, שיכולה להיות מבוצעת בכתב או בעל פה (ראו סקירתו המקיפה של צ'ייף במאמרו המשותף עם טאנן: Chafe & Tannen 1987, וגם הפניות אצל בורכובסקי תש"ע: 8, ואצל רביד תשס"ב: 95, 99–100). על השאלה הזאת אין בכוחנו להשיב כעת, שכן היא מצריכה עוד שנות מחקר מקיפות בתחומי לשון שונים, וסביר שגם אז לא נגיע לתשובה אחת ניצחת אלא לכמה גישות מחקר (ראו למשל Barlow 2011). בהיבטי חקירה רבים אפשר להתנחם בעובדה שחוקרי הלשון העברית חשופים ברובם למחקר לשונות אחרות בעולם, בעיקר הלשון האנגלית, ולקוות שבת קולה של המהפכה העתידית תתגלגל גם למחוזותינו. מכל מקום, ברי שעמדות באשר לגיבוש דרכי חקירה לא יוכלו לנבוע מעמדה תאורטית בלבד, אלא מעיסוק אינטנסיבי בטקסטים ומהתנסות מעמיקה בעבודת מחקר בלשנית. עיסוק שכזה, והוא בלבד, יוכל להוליד את המהפכה הבאה או לחלופין למצוא נוסחה להגדרת היחסים שבין דקדוק הלשון הדבורה ומילונה לאחותה הכתובה.

6. סיכום

עמדנו בקצרה על כמה היבטים בבלשנות המאגר של העברית המדוברת בהשוואה להתפתחות התחום בעולם. ראינו שבעוד בבלשנות העולם התפתח דיון ענף, בעיקר משנות החמישים ואילך, בין הבלשנות התאורטית לבין בלשנות המאגר, מחקר העברית עמד במקום שונה לחלוטין, נאבק על הלגיטימיות שבחקירה הראשונית עצמה.

עם הפריחה בעולם של תחום בלשנות המאגר וכינון מאגרים בני מיליוני מילים, החל בהדרגה לחול מפנה גם בחקר העברית, ואולם נקודת הפתיחה הבעייתית ודלות המשאבים העומדים לרשות קהיליית המחקר גם כיום מותירות אותנו הרחק מאחור. אין ספק שבהיעדר מאגר דיגיטלי מתומלל של שפה דבורה, מחקר הלשון העברית אינו יכול להתקדם בתחומים רבים. שאלות מחקר רבות אינן יכולות כלל להיבדק, ואלו הנבדקות נחקרות באופן שלא תמיד יימצא מְייצג ומקיף. דומה שאין עוררין על כך שכינון מאגר דבור צריך לעמוד בראש סדר העדיפויות של בלשני העברית ולהיות יעד מרכזי לפיתוח המחקר.

ובינתיים אל לנו למשוך את ידינו מחקירת הטקסט האותנטי, אפילו יהא מצומצם ולא מייצג, שכן בשאלות מחקר רבות דגימת נתונים קטנה מספיקה להוכיח הנחות בלשניות, וגם קורפוסים קטנים עשויים להספיק כדי לתאר רבים ממאפייני הדקדוק השכיחים (Ghadessy, Henry & Roseberry 2001, וראו גם את דיונו של Crystal 1997 בלשון הילדים, עמ׳ 231).

ואסיים בנקודת אור אחת אחרונה: למרות היתרון הגדול הגלום במאגרים רחבי היקף, יש לזכור שלפעמים מסד נתונים גדול מאוד גובה מחיר יקר מן החוקר, והחומרים הרבים שנאספו הופכים ל״מטרד מפתה״, שיכול להזיק למחקר או למצער לפגוע בהשגת יעדיו (Miles 1979: 590). ובאמת מתעוררת השאלה: כשעומד לרשות החוקרים מאגר בן כמה מיליוני מילים, עד כמה הם יכולים להכיר את הטקסט באופן בלתי אמצעי, להאזין לו ולא רק להסתמך על תמלילים, וברוח טוניני־בונלי, לתת לטקסט עצמו להוביל את מחקרם. במובן הזה חוקרי העברית החדשה יכולים ליהנות מנחמה פורתא שהיעדר מאגר גדול מזמן להם, ולהעמיק את המחקר הבראשיתי מונע המאגר בטרם יכונַן מאגר גדול ויתגבשו דפוסי חקירה חדשים ומחייבים. הניסיון במחקר עד כה מעלה שגם על יסוד מאגר קטן של טקסטים יכולות להיכתב עבודות מחקר מקיפות וראויות בכל תחומי הלשון, כל עוד נשאלת שאלת מחקר המתאימה לגודל המאגר ולאופיו.

נספח: מקבץ מאגרי לשון חשובים לשפה האנגלית[13]

The Brown Corpus – המאגר הממוחשב הראשון, היה השראה למאגרים נוספים. פותח באוניברסיטת בראון שבארה"ב ע"י Nelson Francis & Henry Kučera, ומכיל למעלה ממיליון מילה של פרוזה אמריקאית כתובה משנת 1961.

The Lancaster-Oslo/Bergen Corpus (LOB) – מאגר בן מיליון מילה משנת 1961 של אנגלית בריטית כתובה. מובא גם בגרסה מתויגת, בחירה אקראית של טקסטים, מקביל לקורפוס בראון, אבל באנגלית בריטית. המאגר נוצר בשיתוף פעולה בין אוניברסיטת לאנקסטר, אוסלו וברגן, והחוקרים המובילים בו: Geoffrey Leech, Stig Johansson.

The International Corpus of English (ICE) – מאגר דינמי הפרוס על פני ארצות שונות דוברות אנגלית. כל תת־מאגר (בארץ נבדקת) כולל כמיליון מילה של אנגלית כתובה ודבורה מארצות שונות ובהן: קנדה, מזרח אפריקה, בריטניה, הונג קונג, הודו, אירלנד, ג׳מייקה, ניו־זילנד, הפיליפינים, סינגפור. החומר הראשוני שנאסף הכיל טקסטים מהשנים 1990–1996, והוא מתויג ומאפשר חיפוש מתוחכם. נוצר על ידי Gerald Nelson מאוניברסיטת הונג־קונג. היחידה הבריטית מנוהלת על ידי The Survey of English Usage.

Michigan Corpus of Academic Spoken English (MICASE) – מאגר דינמי ופתוח לציבור של אנגלית אקדמית דבורה מחיי אוניברסיטת מישיגן, אן־ארבור, ארה"ב. המאגר כולל 1.8 מיליון מילה (מעל 190 שעות) מהשנים 1997–2002. חוקרים מובילים: Römer, Simpson, Briggs, Ovens, Swales.

The London-Lund Corpus of Spoken English (LLC) – מאגר בן חצי מיליון מילה של אנגלית בריטית מדוברת מהשנים 1953–1990. המאגר פותח בשיתוף University College London ואוניברסיטת Lund, שוודיה. יש בו תעתיק מוקפד הכולל סימנים פרוזודיים. חוקרים מובילים: Jan Svartvik, Randolph Quirk, Sidney Greenbaum, Knut Hofland.

13 החומר רוכז באמצעות מקורות שונים, ובראשם פרויקט CoRD: http://www.helsinki.fi/varieng/CoRD/corpora/corpusfinder/index.html ואתרי המאגרים עצמם. לעתים הסתייעתי בפרסומים נוספים הנזכרים בביבליוגרפיה. לא הובאו כאן מראי המקום המדויקים של כל האתרים, שכן אפשר למצאם בנקל בחיפוש במרשתת.

Collins Birmingham University International Language Database (COBUILD, מבוסס על המאגר Bank of English) – מאגר דינמי בן יותר מ־550 מיליון מילה של אנגלית בריטית וגם אנגלית ממקומות גאוגרפיים אחרים, בעיקר אנגלית כתובה אך גם מדוברת. רובו נאסף בשנים 2001–2005. פותח באוניברסיטת בירמנגהאם, אנגליה, בתמיכת ההוצאה לאור Collins. חוקר מוביל: ג׳ון סינקליר.

The British National Corpus (BNC) – מאגר מתויג בן כ־100 מיליון מילה של אנגלית בריטית כתובה (90%) ודבורה (10%) מהשנים 1975–1993.

Corpus The Oxford English – מאגר בן 2.5 מיליארד מילה של אנגלית כתובה מרחבי העולם המבוסס בעיקרו על חומרים מהמרשתת.

Cambridge English Corpus – מאגר המכיל כמה מיליארדי מילים הלקוחות ממקורות שונים: עיתונים, חומרים מהמרשתת, ספרים, רדיו, חומרי לימוד מבתי ספר ואוניברסיטאות – ועוד. הקורפוס אינו פתוח לציבור הרחב.

The Open American National Corpus (OANC) – מאגר בן 15 מיליון מילה של אנגלית אמריקאית כתובה ודבורה משנת 1990 ואילך. הטקסט מתויג ופתוח לציבור.[14]

The Corpus of Global Web-Based English (GloWbE) – מאגר גדול (בתשלום) המכיל 1.9 מיליארד מילה הלקוחות מתוך המרשתת בעשרים ארצות שונות דוברות אנגלית. המאגר נוצר על ידי Mark Davies, פרופסור באוניברסיטת Brigham Young, ופורסם בשנת 2013.

Santa Barbara Corpus of Spoken American Englis (SBCSAE) – חלקו הראשון של הקורפוס פורסם בשנת 2000, והוא כולל הקלטות של אנגלית אמריקאית בדיבור חופשי. חוקרים מובילים: John W. Du Bois, Wallace Chafe, Charles Meyer, Sandra Thompson.

ביבליוגרפיה

באואר מ״ו וארטס ג׳. תשע״א. ״בניית קורפוס: עיקרון מנחה לאיסוף נתונים איכותניים״. בתוך: מ״ו באואר וג׳ גאסקל (עורכים). **מחקר איכותני: שיטות לניתוח טקסט, תמונה וצליל**. רעננה: האוניברסיטה הפתוחה (מהדורה פנימית). 27–49.

בורוכובסקי בר־אבא א׳. תש״ע. **העברית המדוברת: פרקים במחקרה, בתחבירה ובדרכי הבעתה**. ירושלים.

14 המאגר מבוסס על מאגר American National Corpus: http://www.anc.org.

בלנק ח׳. תשי״ז. ״קטע של דיבור עברי ישראלי״. **לשוננו כא**. 33–39.
בן־חיים ז׳. תשי״ג. ״לשון עתיקה במציאות חדשה״. **לשוננו לעם** ד, ג–ה. 3–85.
בן־טולילה י׳. תשמ״ט. ״קורפוס הצרפתית של מונטראול: דגם אפשרי לחקר העברית המדוברת״. **בלשנות עברית** 27. 13–28.
בר־אדון א׳. תשכ״ג. ״לשונו המדוברת של הדור הצעיר בישראל כנושא למחקר״. **החינוך** לה. 21–35.
בר־אשר מ׳. תשס״ג. ״על פועלו של פרופ׳ זאב בן־חיים״. **לשוננו** סה, ג–ד. 227–238.
בר־אשר מ׳. תשס״ט–תש״ע. ״על ריבוי פניה של העברית בת ימינו״. **העברית** נח, א–ב. 5–26.
גונן ע׳. תשע״ג. ״לשאלת הייצוג במחקרים בעברית המדוברת״. בתוך: מ׳ מוצ׳ניק וצ׳ סדן (עורכים). **מחקרים בעברית החדשה ובלשונות היהודים מוגשים לאורה (רודריג) שורצולר**. ירושלים. 417–434.
גושן־גוטשטיין מ׳. תש״ם. ״תחביר העברית החדשה : הרהורים על דרך מחקרה״. בתוך: גב״ע צרפתי ואחרים (עורכים). **מחקרים בעברית ובלשונות שמיות מוקדשים לזכרו של פרופ׳ יחזקאל קוטשר**. 189–201.
וינטנר ש׳. תשס״ב. ״בלשנות חישובית עברית: עבר ועתיד״. בתוך: יזרעאל (עורך). 35–64.
ורום ר׳. תשס״ב. ״הערות מתודולוגיות על כינון מאגר העברית המדוברת בישראל״. בתוך: יזרעאל (עורך). 459–477.
ורטהיימר ע׳. תשס״ד. ״בעיות מינוח בבלשנות השמית״. **בלשנות עברית** 53. 57–74.
יזרעאל ש׳ (עורך, בסיוע מ׳ מנדלסון). תשס״ב. **מדברים עברית: לחקר הלשון המדוברת והשונות הלשונית בישראל** (תעודה יח). תל־אביב.
יזרעאל ש׳. תשס״ג–תשס״ד. ״מחקר העברית המדוברת, הצעד הראשון – על רישום הדיבור לצרכי מחקר״. **לשוננו לעם** נד, ב–ג. 106–118.
יזרעאל ש׳. תשס״ה. ״מאגר הלשון ודקדוק המילון, או: האם הגישות החדשות במילונאות מביאות תועלת או גורמות נזק״. בתוך: מ׳ בר־אשר ומ׳ פלורנטין (עורכים). **מחקרים בשומרונות, בעברית ובארמית מוגשים לאברהם טל**. ירושלים. 335–359.
יזרעאל ש׳. תשע״ו. ״מסד הנתונים למיזם ׳חוקרים עברית מדוברת׳״. בתוך: ע׳ גונן (עורכת). **חוקרים עברית מדוברת** (תעודה כז). תל־אביב. 37–49.
יזרעאל ש׳, הרי ב׳ ורהב ג׳. תשס״ב. ״לקראת כינון מאגר העברית המדוברת בישראל״. **לשוננו** סד. 265–287.
כאן ג׳. תשס״ב. ״חקר העברית החדשה״. בתוך: יזרעאל (עורך). 279–297.
משלר י׳. תשס״ט.״מערכת סמני השיח של העברית היום־יומית הדבורה״. **בלשנות עברית** 62–63. 99–129.
סינקליר ג׳. תשס״ב[1]. ״בלשנות המאגר: שאלות העומדות על הפרק״. בתוך: יזרעאל (עורך). 3–14.
סינקליר ג׳. תשס״ב[2]. ״דקדוק המילון: מבט חדש על השפה״. בתוך: יזרעאל (עורך). 15–33.
רביד ד׳. תשס״ב. ״ניצני ההיצג הדבור והתפתחותם ברכישת השיח העיוני״. **בלשנות עברית** 50–51. 95–120.
רוזן ח״ב. תשט״ו. ״דקדוק העברית הישראלית״. **תרביץ** כד. 234–260.

רוזן ח״ב. תשט״ז. **העברית שלנו**. תל אביב.
רשף י׳. תשס״ד. **הזמר העברי בראשיתו: פרק בתולדות העברית החדשה**. ירושלים.
שורצולד (רודריג) א׳. תש״ע. ״מחקרי העברית החדשה – לאן?״. **לשוננו** עב. 321-336.
שורצולד (רודריג) א׳. תשס״ב. **פרקים במורפולוגיה עברית**. כרך א. תל אביב.
שורצולד (רודריג) א׳. תשס״ח. ״מחקרי העברית בת זמננו בעברית בעשרים השנים האחרונות״. **לשוננו** ע. 449-429.

Al-Sulaiti L. & Atwell E. 2006. “The design of a corpus of contemporary Arabic”. *International Journal of Corpus Linguistics* 11. 135-171.

Aarts J. 2002. “Does corpus linguistics exist? Some old and new issues”. In: L. E. Breivik & A. Hasselgren (eds.). *From the Colt’s Mouth… and Others’: Language Corpora Studies in Honour of Anna-Brita Stenström*. Amsterdam. 1-19.

Baker P. 2010. *Sociolinguistics and Corpus Linguistic*. Edinburgh.

Barlow M. 2011. “Corpus linguistics and theoretical linguistics”. *International Journal of Corpus Linguistics* 16, 1. 3-44.

Barthes R. 1967. *Elements of Semiology*. London.

Bergh G., Seppänen A. & Trotta J. 1998. “Language Corpora and the Internet: A joint linguistic resource”. In: A. Renouf (ed.). *Explorations in Corpus Linguistics*. Amsterdam/Atlanta. 41-54.

Biber D., Conrad S. & Reppen R. 1998. *Corpus Linguistics: Investigating Language Structure and Use*. Cambridge.

Blanc H. 1964. “Israeli Hebrew texts”. In: H. B. Rosén (ed.). *Studies in Egyptology and Linguistics in Honour of H. J. Polotsky*. Jerusalem. 132-152.

Čermák F. 2009. “Spoken corpora design: Their constitutive parameters”. *International Journal of Corpus Linguistics* 14. 113-123.

Chafe W. & Tannen D. 1987. “The relation between written and spoken language”. *Annual Review of Anthropology* 16. 383-407.

Chafe W. 1987. “Cognitive constraints on information flow”. In: R. Tomlin (ed). *Coherence and Grounding in Discourse*. Amsterdam/Phildelphia. 21-51.

Cook G. 1995. “Theoretical issues: Transcribing the untranscribable”. In: G. Leech, G. Myers & J. Thomas (eds.). *Spoken English on Computer: Transcription, Mark-up, and Application*. New York. 35-53.

Crestie M. & Moneglia M. (eds.). 2005. *C-ORAL-ROM. Integrated Reference Corpus for Spoken Romance Language*. Amsterdam/Philadelphia.

Crystal D. 1997. *The Cambridge Encyclopedia of Language*. New York.

De Monnink I. 2000. “A moving phrase: A multi method approach to the mobility of constituents in the English noun phrase”. In: J. Kirk (ed.) *Corpora Galore: Analyses and Techniques in Describing English*. Amsterdam. 133-147.

Fillmore C. J. 1992. "'Corpus Linguistics' or 'computer-aided Armchair Linguistics'". In: J. Svartvik (ed.). *Directions in Corpus Linguistics*. Berlin/New York. 35-60.

Fishman J. A. 1974. "Introduction: The sociology of language in Israel". *International Journal of the Sociology of Language* 1. 9-13.

Gale W. & Church K. 1994. "What is wrong with adding one?". In: N. Oostdijk & P. de Haan (eds.). *Corpus-Based Research into Language – in Honour of Jan Aarts*. Amsterdam/Atlanta.189-198.

Ghadessy M., Henry A. & Roseberry R. L. (eds.). 2001. *Small Corpus Studies and ELT: Theory and Practice*. Philadelphia.

Greenbaum S. 1984. "Corpus analysis and elicitation tests". In: J. Aartsand & W. Meijs (eds.). *Corpus Linguistics: Recent Developments in the Use of Computer Corpora in English Language Research* 45. Amsterdam. 193-201.

Gries S. T. 2010. "Corpus linguistics and theoretical linguistics: A love-hate relationship? Not necessarily…". *International Journal of Corpus Linguistics (IJCL)* 15, 3. 327-343.

Hardie A. & McEnery T. 2010. "On two traditions in corpus linguistics, and what they have in common". *International Journal of Corpus Linguistics* 15, 3. 384-394.

Johansson S. 1995. "Icame - Quo vadis? Reflections on the use of computer corpora in linguistics". *Computers and the Humanities* 28. 252-243.

Joseph B. D. 2008. "The editor's department: Last scene of all… " . *Language* 84. 686-690.

Kennedy G. 1998. *An Introduction to Corpus Linguistics*. London.

Kuzar R. 2001. *Hebrew and Zionism: A Discourse Analytic Cultural Study*. Berlin/ New York.

Leech G. 1991. "The state of the art in corpus linguistics". In: K. Aijmer & B. Altenberg (eds.). *English Corpus Linguistics: Studies in Honour of Jan Svartvik*. London. 8-29.

Leech G. 2000. "Grammars of Spoken English: New outcomes of corpus-oriented research". *Language Learning* 50. 675-724.

McCarthy M. & Carter R. 2001. "Size isn't everything: Spoken English, corpus, and the classroom". *TESOL Quarterly* 35. 337-340.

McEnery T. & Wilson A. 1996. *Corpus Linguistics*. Amsterdam.

McEnery T. & Xiao R. 2011. "What corpora can offer in language teaching and learning". *Handbook of Research in Second Language Teaching and Learning: Volume 2*. 2011. Available at: www.lancs.ac.uk/postgrad/.../Corpora%20and%20 language%20teachingv7.rtf.

Meyer C. F. 2009. "In the profession: The 'Empirical Tradition'". *Linguistics* 37. 208-213.

Miles M. 1979. “Qualitative data as an attractive nuisance: The problem of analysis”. *Administrative Science Quarterly* 24, 4. 590-601.

Ochs E. 1979. “Transcription as theory”. In: E. Ochs & B. Schieffelin (eds.) *Developmental Pragmatics*. New York. 43-72.

Parodi G. 2010. “Research challenges for corpus cross-linguistics and multimodal texts”. *Information Design Journal* 18, 1. 69-73.

Pineda L. A. *et al*. 2010. “The corpus DIMEx100: Transcription and evaluation”. *Language Resources and Evaluation* 44, 4. 347-370.

Quirk R. & Svartvik J. 1979. “A corpus of Modern English”. In: H. Bergenholtz & B. Schaeder (eds.). *Empirische Textwissenschaft: Aufbau u. Auswertung von Text-Corpora*. Königstein/Ts.: Scriptor. 204-218.

Rosén H. B. 1977. *Contemporary Hebrew*. The Hague.

Sankoff D. & Sankoff G. 1973. “Sample survey methods and computer-assisted analysis in the study of grammatical variation”. In: R. Darnell (ed.). *Canadian Languages in their Social Context Edmonton: Linguistic Research Incorporated*. 7-64.

Schmied J. 1993. “Qualitative and quantitative research approaches to English relative constructions”. In: C. Souter & E. Atwell (eds.). *Corpus Based Computational Linguistics*. Amsterdam. 85-96.

Sharoff S. 2006. “Methods and tools for development of the Russian Reference Corpus”. *Language and Computers* 56, 1. 167-180.

Sinclair J. 1985. “Selected issues”. In: R. Quirk & H. G. Widdowson (eds.). *English in the World*. Cambridge. 248-254.

Sinclair J. 2004. *Trust the Text: Language, Corpus and Discourse*. London and New York.

Teubert W. 2010. “My brave old world”. *International Journal of Corpus Linguistics* 15. 395-399.

Thompson G. & Huntson T. 2006. “System and corpus: Two traditions with common ground”. In: G. Thompson & S. Hunston (eds.). *System and Corpus: Exploring Connections*. London. 1-14.

Tognini-Bonelli E. 2001. *Corpus Linguistics at Work*. Amsterdam.

Weiman R. W. 1950. *Native and Foreign Elements in a Language; a Study in General Linguistics Applied to Modern Hebrew*. Philadelphia.

Weinberg W. 1966. “Spoken Israeli Hebrew: Trends in the departures from classical phonology”. *Journal of Semitic Studies* 11. 40-68.

Wilks Y. 2010. “Corpus linguistics and computational linguistics”. *International Journal of Corpus Linguistics* 15, 3. 408-411.

Wilson A., Rayson P. & Archer D. (eds.). 2006. *Corpus Linguistics Around the World*. Rodopi, Amsterdam.