בלשנות המאגר: שאלות העומדות על הפרק

© ג'ון סינקליר, 2000

כשהתבוננתי על השאלות העומדות על הפרק בתחומי בלשנות המאגר (Corpus Linguistics), היססתי מעט אם רצוי שאומר דברים לעניין זה. הנושא אמנם מרתק. זהו ודאי התחום המעניין ביותר והמתקדם במהירות הרבה ביותר מבין ענפיה השונים של הבלשנות. אני אישית מאוד מרוצה מעבודתי בתחום. הרי בשלושים השנים הראשונות שביליתי בעבודתי בבלשנות המאגר לא רבים התעניינו בה, ולכן נעים לראות כה הרבה עמיתים בעלי עניין משותף העוסקים כיום במיגוון רחב של נושאי מחקר הקשורים במאגרי לשון.

הסיבה להיסוסי הייתה ההכרה עד כמה חשוב לארגן את מחקר המאגר על בסיס עקרונות שיאפשרו להגיע לתוצאות ראויות, ולא להיווכח בסופו של דבר שהשימוש במאגרים לא הוביל אלא לתוצאות צפויות מראש. יש תחושה מובנת, במיוחד בקרב קהיליית העוסקים ב"עיבוד שפות טבעיות" (Natural Language Processing או NLP), שכבר ידוע הרבה על השפה, ושהגיעה העת לקצור את פירות המחקר. הלוא בעשרות השנים האחרונות התפתח מחקר תיאורי ויישומי שעשה שימוש במאגרים כעזרים רבי־עוצמה בפרוייקטים שונים. יש טכנולוגיה מתוחכמת, עסקו רבות בבעיות הידועות של המחקר הטקסטואלי כגון עמימות ואנאפורה, וניכרת התקדמות רבה ביישומים עיקריים כגון תרגום ממוחשב.

לצערי, איני יכול להיות שותף להתרוממות הרוח הזו, אף לא אוכל לצפות בשלוות נפש בבקשות הרבות למימון מחקרים היכולות, לכאורה, לאשש התרוממות רוח זו. דעתי האישית היא כי ההישגים בנושא הקשר שבין הלשון והמחשוב מאכזבים, ואינם מבטיחים רבות. בלשנות המאגר מאפשרת להציע כיוונים חדשים בלא המעמסה שנצטברה במשך שנות מחקר רבות. במקום זאת יש עיכובים בהתפתחותה בגלל סגירות מחשבתית שאינה נוטה להיפתח לאפשרויות שיש לה להציע. מכיוון שכך, אתחיל באזכור כמה נקודות מאכזבות

במצב המחקר הנוכחי, ואחר כך אמשיך ואגע בנקודות האור. אולם אל נא תטעו: מאגר לשון אינו משמש כשלעצמו קרש הצלה לאיש; סיכוייו להירתם למחקר גרוע רבים יותר.

טיפוח תקוות

חוששני שכל ההצהרות הראוותניות המושמעות זה שנים בדבר השימוש במחשבים בחקר הלשון הן עדיין בגדר חזון רחוק. לשם המחשת העניין אביא דוגמה אחת. רק לפני שנתיים, בפרסום רשמי של נציבות הקהילייה האירופית, נאמר שבעתיד הקרוב — אולי אפילו נאמר שם "בשנה הבאה" (כלומר, שזה היה צריך להתממש כבר ב־1999) — יעמוד לרשותנו טלפון שנוכל לדבר אל תוכו בשפה אחת, ובן שיחנו ישמע את המסר בשפה אחרת. ככל הידוע לי, הצהרה כזאת הופרחה לאוויר העולם לראשונה באמצע שנות השישים על ידי מרצה בסדרת Reith הידועה של ה־BBC. הצהרות ברוח זו נשמעות גם היום, ארבעים שנה מאוחר יותר. אולם לאמיתו של דבר עדיין רחוקים אנו מאוד מיצירתו של מכשיר כגון זה, וזאת על אף שבתחום זיהוי הקול חלה התקדמות משמעותית.

ככלל, הלשון הוכיחה עד כה שאינה ניתנת לפיצוח. למעשה, לא ניתן לבצע אוטומטית פעולות יעילות על טקסט פתוח. באומרי "טקסט פתוח" (open text) כוונתי לטקסט לא מוגבל, כל טקסט שעשוי לשמש דוגמה סבירה ללשון כלשהי שבשימוש. בעצם, גישות המחקר שנתפתחו לחקר הלשון אינן מטפלות בטקסט פתוח אלא לעתים רחוקות. אשוב לנקודה זו בהמשך, משום שהיא תהווה בסיס לטיעוניי.

קהיליית המחקר פיתחה מבחר מרשים של תהליכי עיבוד שאפשר ליישמם על טקסט שנבחר בקפידה, בהם סוגי תיוג רבים, התורמים "ערך מוסף" לניתוח בלשני ומיגוון כלים סטטיסטיים. אולם אין ביניהם עבודות שיש בהן תועלת, כלומר יישומים שהם אכן בעלי ערך עבור החברה, כאלה שיכולים לבצע פעולות שיתרמו לאיכות החיים. יש, כמובן, פיתוחים מועילים בתחום עיצוב הטקסט, אולם שפות התיוג והסימון וסוגי העיצוב המגוונים להפליא משועבדים למערכות לא משוכללות ומיושנות, וכל אחת מן השפות הללו דורשת המרה של תהליכים ותשומות בעת השקתן זו לזו.

הרושם שלי הוא שהנושאים העיקריים שהעסיקו את הבלשנות החישובית בשנים האחרונות הולכים ונזנחים בהדרגה. איש אינו מדבר עוד על תרגום

אוטומטי מלא. למעשה, רק כמה סוגי תמיכה לתרגום נחשבים ישימים במחשב. מאמצים גדולים מאוד הושקעו בעשור האחרון כדי לייצר מילונים שימושיים ממוכנים, שאפשר יהיה ליישמם על טקסט. הולכת ומתגבשת המסקנה שמדובר בפסבדו־שיטה ולא ביוזמה שתוכתר בהצלחה. לא ייתכן מילון הולם־כול. כפי שנתברר מתוך העבודה על מאגרי לשון, הסיבה לכך היא שעיקר המשמעות נגזר מצירופי מילים שנבחרו לשמש יחד בזמן ובמקום שהבחירה הזו מתבצעת.[1] לגבי לשון אנושית, האוצרת בתוכה אלפים רבים של מילים, המסקנה שהמשמעות נגזרת מצירופי מילים ולא ממילים בודדות, היא "חדשות רעות", משום שמספר האפשרויות מוכפל ללא הכר (אולם ראו עוד להלן). ההשפעה של טיעון זה על יצרני המילונים היא כזאת, שבעוד שמילונים ממשיכים ומתחברים על־פי גישות שונות ועל סמך הנחות מסוגים שונים, הרי שהולכת וגוברת ההכרה כי ההשקעה בכך אינה משתלמת. השאיפה למילון מקיף, שיכלול איכשהו את המידע הלקסיקלי של שפה מסוימת, אינה שאיפה שאפשר לעמוד בה בתחילת האלף השלישי.

המושג של תת־לשון (sub-language) עבר מן העולם. מושג זה שיקף את הרעיון שניתן למצוא בשפה טבעית מיקטעים רציפים בסדרי גודל משמעותיים של סוג שפה שמאפייניה פשוטים יותר. תמיד התמקדו בניסיונות לקבוע קביעות מדעיות יבשות, ללא כל קריצה או פנייה, ותוך דיוק מְרָבי במינוח. והרי תת־לשון בעלת מאפיינים כאלה מגדילה את הסיכויים שמחשב יהיה מסוגל לעבדה. אולם מלומדים שעסקו בחקר השיח (Hunston, 1993) שמו לב שעבודת אנשי המדע לא הייתה אובייקטיבית או חסרת שיפוט ערכי כמצופה. מורים ללשון הצביעו על העובדה כי דווקא מבעים של הערכה משמשים מפתח להבנת הטקסט, ויחד עם זאת לומדי השפה מתקשים לזהות ולעבד מבעים כאלה.

דומני כי התפיסה של תת־לשונות אכן עברה מן העולם. לעומתה, תפיסת הלשונות המבוקרות משגשגת. לשונות מבוקרות הן לשונות המצויות באמצע הדרך לקראת הפיכתן ללשונות מלאכותיות. אלו הן לשונות ערוכות, מטופלות, ומעובדות כך שהמחשב יוכל להתמודד איתן. לשונות אלה הן יותר מכוּונות למחשב מאשר מכוונות לאדם. היצור האנושי הוא שעושה את רוב העבודה בהפקת הלשון המבוקרת, אולם מתאים את התוצר הלשוני שלו למגבלותיה. על־פי אותו קו של מחשבה, הוצע למנוע בעיות בתרגום אוטומטי על ידי עיבוד ראשוני של טקסט המקור (Somers, 1997).

1 לעניין זה ר' מאמרו השני של סינקליר בכרך זה (העורך).

הנה כי כן, נראה כי צריך לעשות הכול כדי להימנע מטיפול בטקסט פתוח, שהרי הוא הבעיה התלויה ועומדת. לכן אני חש כי התמונה המתקבלת ממצב העניינים כעת היא די מייאשת. ישנן אפשרויות וישנו פוטנציאל רב, אך כרגע איני חושב שניתן להצביע על מיגוון רחב של הצלחות.

שאלות העומדות על הפרק

מהן השאלות העומדות על הפרק, אם כן? אילו הן הסוגיות התלויות ועומדות שבכוחן להסיט אותנו מדרכנו או להובילנו לפריצת דרך? הייתי רוצה לבחון סוגיה אחת או שתיים מתחומי בלשנות המאגר, ולאחר מכן סוגיה אחת או שתיים מעבר לבלשנות המאגר, מתחום רחב יותר, תחום מדעי המידע.

ראשית לכול עומדת שאלת ההיקף. מאגרי העיון (reference corpora)[2] הולכים וגדלים, וחצי מיליארד המילים של ה־Bank of English[3] לא נשמע עוד היקף דמיוני. בפרוייקטים רבים כבר שוקלים לכלול במאגר מיליארד מילים או יותר. ככל שהנטייה לשימוש במאגרי לשון גוברת, גוברת עמה גם ההתנגדות למאגרים קטנים, למאגרים מצומצמים, אפילו למאגרים שתוכננו במכוון כך. אומר כמה מילים לעניין זה בהמשך.

שאלה אחרת היא שאלת דרכי סימון המאגר, האופן שבו אנו ״מעשירים״ מאגר בסימנים מפרשים מסוגים שונים. עם זאת עומדת גם השאלה כיצד אנו מכוונים את הניסיונות לקבוע תקנים למבנה המסמך ולעיצובו, כך שכולנו נוכל לתעד ולהפיק מידע על־פי אמות מידה משותפות. אלה הן שתי הסוגיות בתחומי בלשנות המאגר שהייתי מבקש לבחון.

מעבר לתחומה של בלשנות המאגר, אעיר על ההתייחסות אל הלשון כאל מידע. טכנולוגיית המידע היא, כידוע, אחד מתחומי העיסוק החשובים והרגישים ביותר בעולם כיום, והוא מתפתח במהירות. השאלות העומדות על הפרק הן שתיים: כיצד טכנולוגיית המידע מתייחסת אל הלשון וכיצד היא אמורה להתייחס אליה. נראה כאילו אין נדרש מבנה מסוים או אופי מסוים של הלשון כדי לטפל במסמכים כבמידע, ולמעשה הדרישה למבנה כזה נראית מיותרת. ברצוני לבחון זאת בקצרה.

לבסוף, הסוגיה הרצינית, השאלה כבדת המשקל: האם ניתן לתכנת מחשבים כך שיבינו שפות טבעיות? במקום לנסות להשיב על השאלה במעשה עצמו,

2 מאגר עיון הוא מאגר שתוכנן כך שיאצור מידע מקיף ככל האפשר על הלשון.
3 הש׳ הערה 11 במאמרו השני של סינקליר (העורך).

אני סבור כי ישנם שלבים מוקדמים שבהם ניתן לבחון את טיב הבעיה ולבדוק האם היא בכלל ניתנת לטיפול. אם בת טיפול היא השאלה הזו, הרי שאיננו מבינים אותה בדרך שתוביל לתוצאה. אם איננה בת טיפול, הרי שאין אנו אלא מבזבזים את משאבינו כשאנו פועלים מתוך ההנחה שהיא כזו.

מאגרים קטנים

ראשית דבר בתחומי בלשנות המאגר, מה באשר למאגרים קטנים? האם מאגר קטן כמוהו כמאגר גדול רק בהיקף מצומצם, או שיש משמעות לעובדה שהוא קטן? יש, כמובן, מאגרים שהם קטנים מעצם טבעם, משום שעל־פי מיטב ידיעתנו אין הם יכולים לגדול עוד. קיימים הרבה מאגרים קטנים של לשונות מתות אשר אין להם אלא מספר מוגבל וסופי של טקסטים, אלא אם כן יתגלו פתאום טקסטים נוספים. דוגמה נוספת היא מאגר כתביו של סופר חשוב. גם סופרים פוריים לא יגיעו בכתיבתם אלא לעתים נדירות ליותר מאשר מיליון מילים. והרי מאגר כזה אינו אלא דג רקק בקרב מאגרי הענק של היום.[4] פה ושם יימצא טקסט שיש ויכוח על מקורו, אפשר שיתגלה במקרה כתב יד נוסף של אותו סופר, אך עם כל הרצון הטוב והכוונות העניינית, המאגר לא יגדל מכוח ההחלטה בלבד. יש, אם כן, מאגרים קטנים מטבעם שאין לנו שליטה על גודלם ועלינו לנסות ולהפיק מהם את המירב. אולם כאשר מישהו בא ואומר שאין צורך להרחיב את המאגר שלו, שיש בו מספיק מידע, ושלא נחוץ לו שום נתון נוסף, או־אז אני תמה.

נראה לי כי עלינו לבדוק את ההבדל בין מאגר לבין טקסט, ולבחון את ההנחות ואת הכלים שבעזרתם אנו חוקרים ומנתחים את שניהם. טקסט, גם טקסט ארוך למדי, אפשר ללמוד, אפשר לנתח ניתוח יסודי, כולל וממצה, ואפשר לבדוק את מבנהו הדקדוקי עד תום, אלא אם כן הטקסט ארוך במיוחד או שאנו מוגבלים במשאבים. אפשר לראות את תחילתו, את אמצעו ואת סופו. יש לו מבנה, יש בו ארגון מסוים, יש בו הרמוניה במובן הקלאסי של המילה, הרמוניה שניתן לגלות אותה ולדון בה. אתה, המנתח, שולט בעניינים. תוכל לאתר את כל תופעות הטקסט בדייקנות, והן יוכלו להיות בהישג ידך אפילו אם הטקסט ארוך למדי. אלה הם המאפיינים של הדרך שבה אנו עורכים טקסטים למחקר, ושל הדרך שבה אנו חוקרים אותם למעשה.

4 דוגמה אחת מן האנגלית כדי לשבר את האוזן: אנתוני פאואל נפטר בעת כתיבתו של מאמר זה. פאואל תואר כסופר הפורה ביותר מאז פרוסט, והסדרה המונומנטלית שלו, *A Dance to the Music Times*, מכילה כמיליון מילים.

ובכן, כשאוספים מספר טקסטים כאלה, טקסטים שבכל אחד מהם ניתן לטפל בדרך זו, אפשר להתחיל להתייחס לאוסף הזה כאל מאגר לשון. אולם שינוי זה גורר עמו שינוי הכרחי במתודולוגיה: אי־אפשר להעריך מאגר אופייני בדרך של קריאה ושליטה מבוקרות כמתואר לעיל. התחלה, אמצע וסוף של מאגר הם שרירותיים. סדר הטקסטים במאגר הוא בדרך כלל שרירותי, ועצם העניין שבשלו אנו מכוננים מאגר לשון — בניגוד לאוסף של טקסטים — הוא כדי לבדוק דברים שאי־אפשר לבדוק ישירות משום שהם רחוקים זה מזה, משום שהם שכיחים מדי או נדירים מדי, או משום שיתאפשר לנו לגלותם רק לאחר עיבוד מספרי או כמותי כלשהו.

ההבדל בין מאגר לבין טקסט הוא שמאגר לא בודקים באופן ישיר. תחת זאת משתמשים בכלים לעיון עקיף, דהיינו לשונות תשאול (query languages), קונקורדנציות, בודקי קולוקציות, מנתחים תחביריים וכלים ליישור טקסט. יש היום מיגוון רחב של כלים שאפשר להיעזר בהם. חשוב לתת את הדעת לעובדה כי המשתמש במאגר צריך לברור כלים מן המצאי הקיים, או לבנות תוכניות מחשב משל עצמו כדי להפיק את המידע הנחוץ לו מתוך המאגר. משמעות הדבר היא (א) שעל המשתמש לדעת לנסח, לפחות בדרך ראשונית, את השאלה שלגביה הוא מבקש נתונים; (ב) שנדרשות כאן רמות אחרות של פרשנות מעבר למה שהיו רגילים להפיק כאשר חקרו טקסט.

זהו לדעתי ההבדל המכריע בין טקסט לבין מאגר.[5] ההבדל המהותי אינו בהיקף, אף לא ברכיבי התוכן, כי אם במתודולוגיה, בכלי הניתוח ובגישה הניתוחית אל מאגר הלשון. עקרונית, ניתן להתייחס אל טקסט יחיד גדול מאוד כאל מאגר ולטפל בו בטכניקות מאגר במקום בטכניקות טקסטואליות. כך שבמובן מסוים, גודלו של מאגר אינו רלוונטי במיוחד.

אין כל יתרון בקוצר היריעה. היקף קטן מהווה מגבלה. אם נגיע לתוצאות ראויות תוך כדי שימוש בטכניקות המאגר גם ממאגר קטן, הרי המתודולוגיה שלנו היא ללא דופי, אולם התוצאות שנגיע אליהן יהיו מוגבלות מאוד, וטווח המאפיינים הלשוניים שנוכל לצפות בהם יהיה אף הוא מוגבל. היתרון העיקרי במאגר גדול הוא שלחוקיות שבבסיס הדברים יש סיכויים טובים יותר לבוא לידי ביטוי על אף השונות שעל פני השטח. יש לזכור שקיים גיוון רב במימושי היחידות הלשוניות שנקלטו במאגר. אם יסודות לשוניים דומים חוזרים ונשנים בצורה שונה זו מזו, הרי ככל שיופיעו יותר, יהיה קל יותר לראות את החוקיות,

5 להבחנה מפורטת בין טקסט לבין מאגר ר׳ Tognini Bonelli, 2001.

את היסוד החוזר, מאשר את המאפיין הייחודי המלווה כל שימוש אינדיבידואלי של המילים בטקסט.

סיבה משמעותית אחרת לעבודה עם מאגרים גדולים היא האפשרות לרכז את תשומת הלב לא בהישנותן של מילים בודדות אלא בצירופי מילים החוזרים על עצמם, בפרזיאולוגיה וביחידות שהן גדולות יותר מן הצירוף הבודד. נטייה כזו ניכרת לאחרונה במחקר. למדנו מ־Zipf (1935) שרוב המילים אינן נקרות בתדירות רבה מאוד. אם נקיש מזאת לסבירות הופעתם של צמדי מילים, שלשות מילים וכו׳, קל יהיה להבין כי אכן צריך מאגר גדול מאוד כדי לבדוק בדיקה שיטתית את הפרזיאולוגיה. אתן דוגמה אחת: חיפשתי פעם במאגרים אחדים היקרויות של הצירוף fit into place ״להתאים (במקום)״. תחילה חיפשתי במאגר כללי טוב בן שני מיליון מילה, וגיליתי כי לא הייתה שם אף היקרות אחת של הצירוף הזה. והרי יכלה להיות לפחות היקרות אחת, ורק במקרה לא היו שם היקרויות של הצירוף הזה. האומנם? אם ניקח את סיכויי ההיקרות של fit ״להתאים״ ונכפיל אותם בסיכויי ההיקרות של into ״אל תוך״, ושוב נכפיל אותם בסיכויי ההיקרות של place ״מקום״, הרי שאז, למרות שאלה הן שלוש מילים רגילות למדי באנגלית, סיכויי ההיקרות של המילים הללו נמדדים בשברים. כדי לחשב את סיכויי היקרות שלוש המילים הללו יחד, עלינו להכפיל את השברים, והתוצאה תהיה, בסופו של דבר, סבירות הופעה קלושה ביותר של הצירוף המשולש. משום כך, הכפלתי את המאגר פי עשרה, כך שהיו לי עשרים מיליון מילה. אולם גם עתה לא עלתה אף דוגמה, אף כי היה זה מאגר אמין מאוד. מאגר קטן זה בן עשרים מיליון מילה היה המאגר שעל סמך הנתונים שבו נערך המילון הראשון מבוסס המאגר, ובכל זאת לא הייתה בו ולו דוגמה אחת ל־fit into place. הצירוף fit into place הוא צירוף הגיוני ונורמלי לגמרי, ורק העובדה שהוא מורכב משלוש מילים היא שהופכת את מציאתו במאגר לכל־כך לא סבירה. פניתי, אם כן, למאגר בן מאתיים מיליון מילה, שוב פי עשרה, והפעם קיבלתי שש היקרויות. מנקודת ראות סטטיסטית טהורה, הסבירות שיתקבל הצירוף הזה הייתה עדיין קטנה ביותר, אולם הרי מילים אינן פועלות על־פי חוקי ההסתברות; סיכויי כמה מהן להופיע גבוהים הרבה יותר מן הצפי הסטטיסטי. מה שהיה מעניין במיוחד בהיקרויות אלה של הצירוף המשולש fit into place הוא שצירוף זה גרר עמו בחירה מסוימת מאוד של מילים נוספות בסביבתו. למשל, המילה jigsaw ״פאזל, הרכבה״ הופיעה בכמה מתוכן. והרי המילה jigsaw אינה מילה שכיחה באנגלית, ועל־פי כל אמת מידה סטטיסטית מפליא הוא שהצירוף המשולש שלנו, שמופיע רק שש פעמים

במאתיים מיליון מילים, יופיע בהרבה מהן כשהוא כבול למילה jigsaw.[6] מקרה זה, אף אם מרשים במיוחד, הריהו דוגמה מובהקת לכך שכאשר מצרפים יחד מילים על בסיס תדירות היקרותן, על כל צעד ושעל עולים כללי סדר שלא נחזו כלל מלכתחילה. הבחירה המשותפת של המילים היא אמצעי לכללי ארגונו של טקסט.

סימון ותיוג

אפנה כעת לנושא העיון העקיף, ואעיין בו מעט. במתודולוגיה של בלשנות המאגר, העיון העקיף מהווה שלב הכרחי. יש סוגים שונים של עיון עקיף. הייתי מבקש להזהיר מפני שימוש יתר בסוג מסוים של עיון עקיף, והוא הוספה ידנית של "תגים". מטרת התגים היא לסמן היבטים של עיצוב או ניתוח שאינם ניכרים מן ההתבוננות בטקסט גופו. בהקשרים ידועים יכול התיוג להיות אמצעי מועיל, אולם לשימוש הגורף בתיוג יש כמה חסרונות, שמן הראוי שהעוסקים בבלשנות המאגר ישימו לב אליהם, ואשר בגללם כדאי למצוא לו תחליפים.

מסורת התיוג תוסבר על רקע צמיחתה. ראשיתה, כמדומני, לפני כשלושים וחמש שנה, כאשר ראשוני המחשבים, ובעיקר מערכות ההפעלה והתוכנות הראשונות, עדיין לא יכלו לעבד טקסט כטקסט, ואפשרויות עיצוב הטקסט עדיין לא היו קיימות. ראשוני המאגרים אפילו לא יכלו להבחין בין אות קטנה וגדולה בלשונות אירופה. אולם המחשבים יכלו להכיר תגים, ולשונות סימון ותיוג היוו אמצעי להעברת טקסטים ממחשב אחד למשנהו, והן משמשות בזה עד היום. בהדרגה החלו להשתמש בתגים לסימון עוד ועוד היבטים של הלשון, וביניהם גם היבטי תוכן.

אחת מן הבעיות העיקריות בנושא ה"לשון והמידע" שאדון בה בהמשך היא כי הטכניקות של מדעי המידע מאפשרות טיפול במסמכים מבלי להידרש לניתוח הלשון. זוהי ירושה לא מוצלחת במיוחד של השנים הראשונות של הטיפול במידע. כל עוד סומן הטקסט בתגים, יכול היה המחשב לעבוד עם התגים ולהתעלם מן הלשון עצמה. ואמנם, זה מה שעושים המחשבים, כולם עד אחד, למיטב ידיעתי.[7] מכאן נובע בבירור שהשימוש הבלתי מבוקר בתיוג מהווה

6 בבדיקה מאוחרת יותר במאגר הענק של ה־Bank of English מופיעה המילה jigsaw כבולה לצירוף fit in place רק שלוש מתוך שמונה פעמים שבהן נקרה הצירוף הזה במאגר. עדיין הממצא ייחשב כיוצא דופן על־פי כל אמת מידה סטטיסטית.

7 יש סימנים לכך כי כמה ממנועי החיפוש המתוחכמים וכמה מתוכניות התמיכה במסמכים בכל זאת משתמשים בכמה יסודות לשוניים מובהקים וברורים של משמעות, כגון

תחליף לעיון בלשני: לומדים את התגים ולא את הלשון, וכל מי שמשתמש בתגים ראוי לו שידאג להשתמש באמצעי זה בדרך לעיון הבלשני, אולם לא במקומו.

שיבוץ תגים בטקסט לשוני הוא פעילות מסוכנת, משום שהטקסט מאבד בדרך זו את אחדותו. עד כמה שיהא המשתמש זהיר בעבודתו, לאחר תיוג אי־אפשר עוד לשחזר את הטקסט המקורי בנאמנות. למרבה המזל, היום אין אנו זקוקים עוד לשילוב שתי הרמות, זו של הטקסט וזו של התיוג, אלא אך לעתים רחוקות ורק כשתוכנת המחשב דורשת זאת. ב־Bank of English, למשל, רצפי התגים תמיד מופרדים מגוף הטקסט על ידי הזרמת נתונים מקבילה. אחת מן הבעיות בתיוג שעדיין לא נפתרו היא הצורך המתמיד בהתערבות אנושית. מכיוון שהמודלים האנליטיים שעל־פיהם מתויגים הטקסטים פותחו בידי בן אנוש, אין הם מאפשרים קטגוריזציה מושלמת של הנתונים, ומשום כך אין התוכנות יכולות להפיק ולשמר תוצרים שאפשר להשתמש בהם. כמובן, תוצאה נוחה של השימוש בטקסט מתויג היא כי התיאור שמפיקים התגים אינו בר הפרכה, שכן הוא מוגן על ידי מערכת התיוג עצמה: את נתוני המאגר הלוא אפשר לראות רק דרך משקפי התגים, כלומר: כל מה שהתגים לא יוכלו לסמנו — יחסר. ולבסוף, כתופעת לוואי, הטקסט הופך דחוס בתגים עד סף התפוצצות, ומהירות העיבוד קטֵנה.

כך היא דרך העבודה שבה הקטגוריות לתיאור מאגר מותאמות לתיאור מאגר, ותוצאתה תיאור מוגן של המאגר. אלנה טוניני־בונלי (Elena Tognini-Bonelli, 2001) מכנה דרך זו באופן כללי "בלשנות מבוססת מאגר" (corpus-based linguistics). היא מבדילה את "הבלשנות מבוססת המאגר" מן "הבלשנות מונעת המאגר" (corpus-driven linguistics), וזאת הגישה שגם היא וגם אני נוקטים. הבלשנות מונעת המאגר היא שיטה שאינה משתמשת בטקסט מתויג אלא מעבדת ישירות את הטקסט הגולמי כך שדפוסים הניכרים מתוך הטקסט הגולמי צצים ועולים לעיני החוקר. ניתן לעבד את יחידות הטקסט עצמן, ואני שמח לדווח שעומד לרשותנו מיגוון גדל והולך של תוכנות ואמצעים נוספים, הפועלים ישירות על הטקסט הטבעי ומספקים תוצאות מרשימות. בעתיד הקרוב אני מקווה להעלות על אתר האינטרנט שלנו (www.twc.it) רשימה של תוכנות כאלה, רשימה שתכיל תוכנות הפועלות על קובצי טקסט או על קבצים שנוצרים מתוכנה אחרת ברשימה הזאת. בצורה זו, הגישה לנתונים נשמרת בדרך הפשוטה ביותר, ועיבוד הנתונים וניתוחם הם אוטומטיים לגמרי.

קולוקציות. אולם הם עושים זאת "מבחוץ", אם אפשר להתבטא כך, ובלא מודעות לשילובם עם דגמי לשון אחרים.

עד כאן דנתי בקצרה בשני הנושאים העיקריים בתחום בלשנות המאגר: היקף המאגר והגישה המתודולוגית לתיוג. עתה ברצוני להזכיר בקיצור רב עוד יותר את שאלת מדעי המידע ואת הדרך הבלתי־הולמת שבה מתייחסים כיום מדעי המידע אל הטקסט הלשוני. הטקסט נראה לעוסקים במדעי המידע כלא מובנה וכבעל איכות ירודה מבחינת המידע, ומשום כך הם מתעלמים מדפוסי הטקסט וממבני הלשון ובמקום זה הם מסרבלים אותו בתיוג, שהיא פעולה יקרה. זוהי תוצאה והרחבה של עקרון התיוג, המוחל כאן באופן גורף. אין כלל התייחסות למבנה הלשון בתעשייה הענפה של אחזור מידע. אחת הסיבות העיקריות למצב עניינים עגום זה היא רשימת ההישגים העלובה של ה־NLP (עיבוד שפות טבעיות), מצב הנובע מהסתמכות יתר על תיאוריות שהוכחו כלא מתאימות. על נושאים אלה ועל מה שנוגע להם דנתי במאמרים אחרים (Sinclair, 1999, 2000), ולא אשוב ואדון בהם כאן.

הנושא האחרון שאני מבקש להזכירו נובע מן העניין הקודם, ונוגע לנחיצותה של הבלשנות בעולם הדיגיטלי, בעולם האינטרנט ורשתות התקשורת, ובעולם של מידע בכלל. השאלה היא האם יכולים אנו להניע מחשבים להתנהג כאילו הבינו טקסט פתוח בשפה טבעית. אני מאמין שתחילה עלינו לתת את דעתנו לשאלה, אם אפשר יהיה אי־פעם לסמוך על המחשב שיבין ולו את הטקסט הפתוח הפשוט ביותר. אם נגיע למסקנה שמשימה זו הינה בת ביצוע, או־אז ניגש ליישומה, ועל כל קהילת בלשנות המאגר להתייצב בחזית מאוחדת אחת לתקיפת הבעיה שלב אחרי שלב, משום שהכרעתה תהיה משתלמת מאוד. תארו לעצמכם שתוכלו לתקשר עם המחשב שלכם בלשון של יומיום ולא דרך תפריטים, קליקים ובקשות עזרה פאתטיות. תארו לעצמכם שתוכלו לבטוח בהודעות המחשב, שההודעות ייאמרו בלשון טבעית, ושלא נצטרך לגחך עוד אל מול ההודעות הללו כפי שאנו עושים היום.

אולם אם, לעומת זאת, נגלה, כפי שטוענים בני סמכא רבים, כי יש נושאים מסוימים בלשון הטבעית החסומים בפני המחשב, ושאין הוא מסוגל להתמודד איתם — או־אז יהא עלינו לשקול מחדש את עמדתנו לגבי מה שאנו, כקהיליית מחקר, מכריזים שאנו יכולים לעשות עתה ומה שנוכל לעשות בעתיד. במקרה כזה נצטרך גם לשקול שוב מה נוכל להציע לחברה כולה בעזרת מומחיותנו בטיפול בשפה. יהא עלינו לעשות הערכה מחודשת, ואולי אף להנמיך ציפיות.

לסיכום, אני מבקש להצביע על העובדה כי ארבעת הנושאים שהעליתי קשורים זה בזה באופן מורכב למדי. אם לא תהיה לנו גישה למאגרי לשון עצומים בגודלם, לא תהיה לנו גישה למידע הלשוני הנחוץ לנו כדי לעמוד

באתגרים שמציבה בפנינו הדרישה לאחזור מידע. ככל שאנו ממשיכים להסתמך על תגים, תשומת לבנו (והמשאבים העומדים לרשותנו) מופנית בהכרח אל עבר מודלים שפותחו בתקופות של טרם מאגר, מודלים שממילא דורשים אך מאגרים קטנים. מאגרים מתויגים לא יעמדו בדרישות חברת המידע, משום שאינם רגישים דיים. לו היו, היו כבר מנצחים בתחרות, משום שזכו לכל תשומת הלב הנדרשת עד עתה. מאגרים אלה הוכחו ככושלים במיוחד בטיפולם בטקסט פתוח, שהוא עיקר הטקסט שיידרש לטיפול ממוחשב בהבנת הלשון האנושית.

לבלשנות מונעת המאגר נדרשים מאגרי לשון עצומים בגודלם, משום שהיא דורשת היקרויות רבות ככל האפשר של היחידות שהיא מטפלת בהן. סוג בלשנות זה אינו מתיר שימוש בתיוג ידני ומזמין חשיבה מחודשת על הדרכים לניתוח ממוחשב. הוא פותח דרכים חדשות למחקר, שעשויות לעזור לאחזור מידע וליישומים אחרים, והוא עשוי לקרבנו אל המטרה של הבנת הלשון האנושית על ידי המחשב. אין צריך לומר שהמתנגדים לאפשרות זו חייבים להעמידה בניסיונות רבים לפני שיתקפוה.

ביבליוגרפיה

Bank of English, The. <www.cobuild.collins.co.uk>

BNC Sampler. 1999, release 1.1 (CD-ROM): Oxford University Humanities Computing Unit.

Hunston, S. 1993. Evaluation and Ideology in Scientific Writing. In: M. Ghadessy (ed.) *Register Analysis: Theory and Practice*. London: Pinter. 57–74.

Sinclair, John. 1999. New Roles for Language Centres: the Mayonnaise Problem. In: D. Bickerton and M. Gotti (eds.). *Language Centres: Integration through Innovation*. CercleS (Confédération Europèenne des Centres de Langues de l'Enseignement Superieur). Secretariat, Department of Modern Languages, University of Plymouth. 31–50.

Sinclair, John. 2000. The Deification of Information. In: G. Thompson and M. Scott (eds.). *Patterns of Text: in Honour of Michael Hoey*. Amsterdam: John Benjamins. 287–314.

Somers, H. 1997. A Practical Approach to Using Machine Translation Software: "Post-editing" the Source Text. *The Translator* 3: 193–212.

Tognini Bonelli, Elena. 2001. *Corpus Linguistics at Work.* (Studies in Corpus Linguistics, 6.) Amsterdam: John Benjamins.

Zipf, G. K. 1935. *The Psychobiology of Language*. Cambridge, MA: MIT Press. (Reprinted: 1965.)

פרסומים חשובים נוספים של ג'ון סינקליר הנוגעים בנושאים הנידונים במאמר זה:

1991. *Corpus, Concordance, Collocation.* (Describing English Language.) Oxford: Oxford University Press.

1995. From Theory to Practice. In: G. Leech, G. Myers and J.Thomas (eds.). *Spoken English on Computer*. London: Longman.

1995. Computers and Language Teaching. In: I. Lee et al. (eds.). *Selected Papers from SICOL-1992*. Seoul: Hanshin Publishing Company. 287–297. (= A. Wichman et al., eds. *Teaching and Language Corpora*. London: Longman, 1997. 27–39)

1997. Corpus Linguistics at the Millennium. In: J. Kohn et al. (eds.). *New Horizons in CALL*. Szombathely: Bersenyi Dániel College.

2000. The Computer, the Corpus and the Theory of Language. In: Gabriele Azzaro and Margherita Ulrych (eds.). *Transiti linguistici e culturali*. Volume II: Proceedings of the XVIII AIA Congress Anglistica e...: *metodi e persorsi comparatistici delle lingue, culture e letterature di origine europea*. Trieste: EUT. 1–15. (=*LMS Lingua* 1/99: 24–32.)