

הדגמת יכולות אחזור המידע בעברית במארג (WEB) החופשי תמונת מצב לתחילת 2005

יהודית בר-אילן

מאמר זה בוחן את טיב אחזור המידע בעברית מה-Web של מספר כלי חיפוש פופולריים. הבחינה נעשתה בעזרת שאלות המדגישות קשיים מיוחדים של השפה. תוצאות הבדיקה מראות שבתחילת 2005 הכלים שנבדקו אינם מעודכנים ואינם יכולים להקיף את הבעיות המיוחדות של השפה העברית או לטפל בהן, בעיקר בגלל מורכבותה המורפולוגית.

מבוא

לפי מילון המונחים של האקדמיה ללשון (2004) יש לקרוא ל-Web בשם המארג או מארג כלל עולמי. לא ברור האם מונח זה יתפוס, כי גם עבור ה-Internet הוצעה מלה, מרשתת, אך במילון המונחים המעודכן כבר מופיעה המלה אינטרנט כמלה חוקית. אי לכך נשתמש הן במונח Web והן במונח "המארג" במאמר זה. אמנם ה-Web קיים כבר 15 שנה, אך בתחילת הדרך הופיע רוב המידע שבו בשפה האנגלית. ביחד עם זאת, אחד משרתי ה-Web הראשונים הוקם באוניברסיטה העברית בירושלים כבר בשנת 1992 (Berners-Lee, 1992). בשרת הזה ניתן היה לדלות חומר באמצעות דפדפן טקסטואלי בעל ממשק עברי וכן להציג מידע בשפה העברית, וזאת למרות מיעוט יחסי של חומרים בעברית בשלב הזה. לפי האתר "דפי רשת" (Dapey Reshet, 1999)

בתחילת 1997 היו בסך הכל 300 אתרים (sites) שתמכו בהצגת מידע בעברית - כ-30 אחוז מכלל האתרים הישראליים שהיו קיימים אז), ובסוף שנת 1997 מספרם גדל לכ-2,500 (כ-55 אחוז מכלל האתרים הישראליים שהיו קיימים אז). בתחום il (domain) ובשפה העברית, ממפתח מנוע החיפוש Google נכון לסוף נואר 2005, כ-4,140,000 דפים (pages), המהווים כ-43.4 אחוז מכלל הדפים ש-Google ממפתח בתחום il. כמובן שקיימים דפי Web נוספים שאינם בתחום il, שאף הם כתובים בשפה העברית. כך, שלכל הדעות מדובר בכמות מידע בלתי מבוטלת. יש לשים לב לעובדה שבספירה של "דפי רשת" נספרו אתרים, והמידע העדכני מתייחס למספר הדפים, כך שלא ניתן להשוות באופן ישיר בין המידע ההיסטורי למידע העדכני. ביחד עם זאת ברור, שחל גידול עצום בכמות המידע בעברית הנמצאת ב-Web.

איתור תכנים ב-Web מתאפשר במספר דרכים:

עיתונות, תקשורת - למשל המדור השבועי "עלו ברשת" של "ידיעות אחרונות" או בפרסומות ברדיו ובטלוויזיה.

חברים, קולגות - בעל פה או באמצעות דואר אלקטרוני. אנשים נוטים לסמוך על המלצות חבריהם גם בתחום זה. רשימות של אתרים בנושא מסוים שהוכנו על ידי מומחים - לדוגמה, רשימת המשאבים של ה-Public (PLA - Library Association) <http://www.pla.org/PLAtemplate.cfm?section=resources> (כמו במקרה של פנייה לאתר מסוים בעקבות המלצות חברים, גם פה מייחסים בדרך כלל חשיבות למקור המידע: ככל שהמקור סמכותי יותר, כך תגבר הנטייה לקבל את ההמלצה).

מדריכים בהם האתרים הנסקרים משויכים לסיווג מסוים - המדריך הכללי המוכר ביותר בעולם הוא Yahoo! (<http://www.yahoo.com>). יש כמה מדריכים בעברית, לפי סקר TIM האחרון (טלסקר/TNS, 2004), הפופולארי מביניהם הוא "וואלה!" (<http://www.walla.co.il>), אחריו "נענע" (<http://www.nana.co.il>) ו"תפוז" (<http://www.tapuz.co.il>). אתרים אלה אינם מדריכים בלבד, אלא פורטלים רבי היקף. בדף הבית של "תפוז" אין כלל אזכור לכך שקיים גם מדריך, ורק כאשר המשתמש יפנה אל "אינדקס אתרים" יגיע לדף הפתיחה של המדריך. יש להניח אם כן שעיקר פעולתו של "תפוז" אינו בתחום סיווג מידע. ביחד עם זאת, ממצאי סקר TIM מנובמבר 2004 (ynet, 2004) עולה, שהפעולה הנפוצה ביותר בקרב משתמשי האינטרנט היא חיפוש מידע (93.5 אחוז מכלל הגולשים עוסקים בחיפושים).

מדריכי האתרים בפורטלים האלה מספקים תיאורים קצרים של אתרים נבחרים בשפה העברית, גם אם התכנים באתר נכתבו בשפה אחרת (בעיקר באנגלית) ובכך מקווים, כנראה, לעזור למשתמשים להתמצא ביתר קלות גם באתרים לועזיים.

כלי חיפוש המחפשים בטקסט הנמצא בדפי ה-Web - אלה מנועי החיפוש. הפופולארי ביותר ברוב העולם הוא Google (Sullivan, 2004). בישראל הוא לא רק כלי החיפוש הפופולארי ביותר, אלא גם האתר שמספר הגולשים המבקרים בו הוא הגדול ביותר: 62 אחוז ממשתמשי ה-Web ביקרו בשבוע שקדם לסקר TIM לפחות פעם אחת ב-Google (טלסקר/TNS, 2004).

בגלל כמויות המידע העצומות הקיימות ב-Web, רק שתי האפשרויות האחרונות, ולמעשה בעיקר האופציה האחרונה, מעשיות כאשר יש צורך בכלי המאפשר איתור מידע ב-Web בנושא כלשהו. לפי מחקר של ה-Pew Internet and American Life Project שהתפרסם בתחילת ינואר 2005 (Fallows, 2005), חיפוש מידע הופכים להיות פעילות יומיומית אצל כשליש ממשתמשי האינטרנט האמריקאיים. פרק הזמן הממוצע של חיפושים אצל הגולש האמריקאי הוא 43 דקות בשבוע (34 שאילתות). גם בסקר זה התברר ש-47 אחוז מהנשאלים שמשתמשים במנועי חיפוש פונים בתכיפות הרבה ביותר ל-Google. לגבי המצב בישראל אין בידינו מידע מפורט כל כך, אך ברור שגם הגולש הישראלי מחפש מידע. אין לנו גם נתונים על שפת החיפוש (עברית, אנגלית או שפה אחרת), אך יש להניח שאחוז בלתי מבוטל מהגולשים ישראלים (וכנראה במיוחד בני הנוער) מחפשים מידע בעברית. לכן, יש חשיבות רבה לבחינת יכולות האחזור של כלי החיפוש ב-Web המסוגלים להשיב לשאילתות אשר נוסחו בשפה העברית.

כפי שנאמר בשירו של דן אלמגור "עברית קשה שפה", ועל אחת כמה וכמה כאשר היא נכתבת במחשב. קיימות מספר שיטות קידוד מקובלות (למשל עברית ויזואלית ועברית לוגית). קושי נוסף מבחינת הצגת המידע הוא כיוון הכתיבה מימין לשמאל (רוזן, 1996). בנוסף לבעיות הטכניות שהוזכרו לעיל, עברית היא שפה מורכבת מבחינה מורפולוגית. השפה האנגלית, לעומת זאת, היא שפה פשוטה מאוד בצורתה ובהטיות לשמות העצם והתואר וגם מילות היחס באנגלית עומדות בפני עצמן ואינן חלק מהמלה, כמו שקורה בדרך כלל בעברית. יש אמנם שפות נוספות בהן הוספת מילת יחס ושייכות משנה את צורת המלה, אך בדרך כלל השינוי חל בסוף המלה ולא בתחילתה.

ניח שמחפשים מידע מקיף על אבוקדו (זה עוד מקרה פשוט כי מלה זו אינה מופיעה ברבים וגם אין משתמשים בסופיות כדי להביע מילות שייכות, כלומר לא אומרים "אבוקדוי" כאשר רוצים לדבר על האבוקדו שלי). מכיוון שהמלה אבוקדו יכולה להופיע בצורות שונות בטקסטים של דפי ה-Web (כגון: האבוקדו, לאבוקדו, מאבוקדו, מהאבוקדו), אם כלי החיפוש אינו מרחיב את החיפוש לצורות המורפולוגיות השונות, יהיה עלינו לחפש את מופעי המלה בכל הצורות (כאשר אנחנו מעוניינים במידע מקיף על הנושא). בניגוד לעברית, באנגלית ניתן להסתפק בחיפוש המונח avocado, כי חיפוש זה יאתר מופעים גם של the avocado, an avocado, from the avocado וכו'.

אתגר נוסף עבור כלי אחזור בשפה העברית היא העובדה שלא נעשה שימוש בניקוד. כתוצאה מכך, רצף אותיות מסוים יכול להיות בעל משמעויות רבות, וזאת בהשוואה לשפות בהן התנועות (vowels) הן חלק מהאותיות ולא ניתן להשמיטן כפי שניתן לעשות באותיות הניקוד בעברית. ניקח לדוגמה את רצף האותיות מספר.

ללא ההקשר אין אנו יכולים לדעת האם מדובר ב-tells או ב-coiffed או ב-number או ב-from a book או ב-from a hair stylist (בדיקה במילון מורפיקס - <http://milon.morfex.co.il/MorDictFirstPage.htm> מעלה אפשרויות נוספות).

לסיכום, רוב המאמצים והמשאבים מושקעים בפיתוח כלי אחזור מידע בשפה האנגלית. והסיבה מובנת. לפי Global Reach (2004) יש בסך הכל 3.8 מיליון גולשים ששפתם העיקרית היא עברית, בעוד שיש 295.4 מיליון משתמשי אינטרנט אשר שפתם העיקרית היא אנגלית, ובנוסף לכך, גם עבור חלק בלתי מבוטל של הגולשים האחרים (מוערכים בכ-544.5 מיליון) משמשת האנגלית שפה שנייה.

למרות שרק חברות מסחריות מעטות משקיעות בפיתוח כלי אחזור בשפה העברית, נעשו עבודות מחקר מעמיקות בתחום זה - מאמרי סקירה מקיפים בנושא נכתבו על ידי Wintner (2004) ועל ידי מרגלית (2004).

ראוי במיוחד לציין את מחקרי החלוץ של שויקה ושותפיו (Attar et al., 1978; Choueka, 2005). מנוע החיפוש מורפיקס (<http://www.morfis.co.il>) משלב טכנולוגיות שפותחו על ידי קבוצה בראשותו של שויקה (מלינגו, 2003).

מערך הבדיקה

מטרת הבדיקה היא לבחון את יכולות האחזור של כלי החיפוש הפופולאריים בשפה העברית ב-Web. אין אנו בוחנים את האלגוריתמים או מנגנוני הפעולה של כלים אלה, אלא את תוצאות פעולתם מנקודת מבטו של המשתמש. מטרתנו היא להסב את תשומת לב הקוראים לאחזור החלקי בעברית. בהתאם לממצאי הסקר האחרון של TIM (TNS/טלסקר, 2004) נבדוק את הכלים הבאים: Google, וואלה! ונענע (<http://www.nana.co.il>).

Google הוא מנוע חיפוש המחפש בטקסט החופשי של דפי ה-Web, לוואלה! יש גם מדריך וגם מנוע חיפוש עבור דפי Web (נבדוק את שניהם, אך בנייתו המדוקדק נתייחס לתוצאות המדריך בלבד). החיפוש בנענע הוא בתכנים של המדריך בלבד. וואלה! ונענע מאחזרים גם כתבות חדשותיות. בנושאים אקטואליים נתייחס גם לתוצאות אלה. בנוסף לכלים אלה נבחן גם את מורפיקס, בהיותו הכלי היחיד המצהיר על "טיפול נכון בבעיות מחשוב שיוצרת המורפולוגיה" (מורפיקס, 2001).

בחרנו שאילתות האמורות להבליט בעיות שונות. החלטנו להתמקד בנקודות הבאות:

- * מידת עדכונן של הכלי וזאת באמצעות השאילתות "תוכנית ההתנתקות" ו"צונאמי", שהיו נושאים אקטואליים בינואר 2005. מאגר שאינו מעודכן לא יניב תוצאות בעלות ערך למשתמשיו גם אם הוא מתמודד מצוין עם הקשיים הלשוניים שהעברית מציבה בפניו.

- * יכולתו להתמודד עם ההבדלים בין כתיב מלא ולכתיב חסר.
- * יכולתו להתמודד עם הוספת תחיליות וסופיות (באמצעות וריאציות שונות של המלה "ספרייה").
- * יכולתו להתמודד עם שימוש בגרשיים לציון קיצור בעברית (באמצעות השאילתות "נוה אטי"ב" ו"ניות").

תוצאות האחזור ינותחו לפי הערכת טיב התוצאות הראשונות. מחקרים מראים שהמשתמש הממוצע ברוב המקרים מסתפק בבחינת התוצאות הראשונות (Spink et al., 2001; Silverstein et al., 1999). אנו נבחן את חמש התוצאות הראשונות עבור כל שאילתה וניתן את חוות דעתנו הסובייקטיבית. בשאילתות הבוחנות את יכולת ההתמודדות עם צורות מורפולוגיות שונות, ציינו גם את מספר התוצאות המאוחרות, כי הן מצביעות על טיפול או על חוסר טיפול בצורות המורפולוגיות השונות. ברור שאין להשוות בין מספר התוצאות המאוחרות על ידי מנוע חיפוש אשר אוסף את התוצאות באמצעות זחלן (crawler) לבין מספר התוצאות במדריך שהמידע שנאסף בו נבדק על ידי עורכים אנושיים וממוין על ידיהם. כך שבמאמר זה יושוו אך ורק מספר התוצאות שמתקבלות מאותו הכלי עבור וריאציות שונות של השאילתה. בשיטה זו השתמשו כבר בר-אילן וגוטמן (Bar-Ilan, & Gutman, 2005) כדי לבחון יכולות אחזור ברוסית, בצרפתית, בהונגרית ובעברית, גוגנהיים ובר אילן (Guggenheim, & Bar-Ilan, 2005) כדי לבחון אחזור בגרמנית ובעברית ו-Moukdad (Moukdad, 2004) כדי לבחון אחזור בערבית.

בכל אחזור בדקנו לעומק חמש תוצאות ראשונות. בדיקה זו מאפשרת לחשב את המדד (Baeza-precision@5) (Yates & Ribeiro-Neto, 1999) שהוא אחוז התוצאות הרלוונטיות מתוך חמש התוצאות הראשונות. למדד זה חסרונות רבים הנובעים בעיקר מסובייקטיביות המושג "רלוונטיות". כדי לצמצם את הבעיה, אפיינו לפני תחילת הבדיקה מה נחשב רלוונטי בעינינו. חיסרון נוסף של המדד הוא בכך שאין אפשרות לציין את מידת הרלוונטיות של התוצאה. אם למשל אנחנו מחפשים אתרי חדשות ישראליים, אנו מצפים לקבל כתשובה אתרים כמו ynet, "הארץ" או NFC ונהיה מאוכזבים אם בתוצאות הראשונות יכללו רק אתרי חדשות שמתחזקים על ידי גופים לא מוכרים או אנשים בלתי ידועים, למרות שגם אתרים אלה עונים על דרישת החיפוש. בדיקה זו זה מדגימה את הבעייתיות של אחזור בשפה העברית באמצעות ניתוח מספר קטן של שאילתות. לא היה צורך לבצע ניתוחים סטטיסטיים, הרי שהכלים (פרט למורפיקס) אינם טוענים שהם מטפלים בבעיות המיוחדות של השפה העברית. כפי שציינו לעיל, מטרת המאמר היא להסב את תשומת לב הקוראים לכך, שהאחזור בעברית הוא חלקי בלבד.

תוצאות

עדכניות הכלים

לצורך בחינת העדכניות בחרנו שתי שאליות:

"תוכנית ההתנתקות" (חיפוש כביטוי בכלים המאפשרים זאת) וחיפוש המלה "צונאמי". בהסתמך על תוצאות החיפוש בארכיון "הארץ" (<http://www.haaretz.co.il>), המושג "תוכנית ההתנתקות" הוזכר בפעם הראשונה ב-10 בדצמבר 2003 ונכנס לשימוש יומיומי החל מסוף דצמבר 2003. בחיפוש מידע עבור "צונאמי" רצינו לבדוק אם יש מידע על הצונאמי ששטף את חופי דרום מזרח אסיה בעקבות רעש אדמה ב-26 בדצמבר 2004.

כל הכלים פרט למורפיקס העלו תוצאות המתייחסות ל"תוכנית ההתנתקות" במובנה העכשווי. המצב היה דומה עבור השאלית "צונאמי". במקרה זה, אחזר מורפיקס 25 תוצאות, אך אף אחת מהן לא התייחסה לאירוע של סוף 2004.

מתוך בדיקות אלה עולה החשד שבתחילת 2005 לא היה מורפיקס מעודכן, אך לפני שמגיעים למסקנות מרחיקות לכת יש צורך לבצע בדיקות רחבות יותר. ביחד עם זאת, כאשר רוצים להשתמש בכלי אחזור Web-ב, רצוי לבצע בדיקות מדגמיות שיכולות לרמוז על עדכניות התוצאות.

תחיליות, סופיות, כתיב מלא וכתיב חסר:

לצורך בחינת יכולות האחזור בשפה העברית בחרנו במספר שאליות הקשורות במונח "ספרייה". מספר התוצאות המדווחות מוצגות בטבלה מספר 1. מטרתן של השאליות אותן הצגנו היתה לדלות מידע בנושא "ספריות" (libraries) בעברית. קישור שאינו פועל נחשב לבלתי רלוונטי, גם אם התיאור הקצר שכלי החיפוש מציג מרמז על כך שהדף עשוי להיות רלוונטי. תוצאה שיש ממנה קישור לספרייה נחשבת לבלתי רלוונטית, אלא אם הדף מקשר למספר גדול של ספריות (רשימת ספריות).

Google

נבחן תחילה את התוצאות שהועלו באמצעות גוגל: כבר מתוך הטבלה ניתן לראות שאין שום טיפול בצורות המורפולוגיות, (יש יותר תוצאות לשאלית "הספרייה" מאשר לשאלית "ספרייה"), ויש מעט מאוד מסמכים (שאלית 4) בהם מופיעות שתי הצורות. כך, שכאשר בוחרים לחפש באחת הצורות בלבד, מפסידים מסמכים. נוגל אינו מטפל בכתיב החסר לעומת הכתיב המלא, וכתוצאה מכך, גם במקרה הזה לא נדלים חלק מהמסמכים הרלוונטיים מכיוון שמספר המסמכים בהם מופיעה גם המלה "ספרייה" וגם המלה "ספרייה" הוא 1,080 בלבד.

בהמשך בדקנו עבור כל שאלית את מידת הרלוונטיות של חמש התשובות הראשונות. **שאלית מספר 1 - "הספרייה":**

התוצאות הראשונות עבור השאלית כוללות רשימת ספריות, אתר של ספרייה וירטואלית לענייני הלכה, אתר הבית של הספרייה של המכון האקדמי-טכנולוגי של חולון, ספרייה דיגיטאלית של מחוז דרום של משרד החינוך (בניגוד לקודמים הדף אינו מכיל את המלה ספרייה, שהיא צורת המלה בכתיב מלא, אלא היא מופיעה בו בכתיב חסר בלבד), ואת דף הבית של הספרייה למשפטים של אוניברסיטת בר-אילן (דף זה אינו מכיל את המלה "ספרייה", ומופיע בו רק המלה אלא מופיעה בו רק המלה "הספרייה" בלבד, וגם המטמון של גוגל מאשר זאת: המלה מופיעה רק בקישורים המובילים לדף). במקרה זה כל התוצאות היו רלוונטיות, לכן precision@5 הוא 100 אחוז.

שאלית מספר 2 - "הספרייה":

במקרה זה שתי התוצאות הראשונות הם דפים מהספרייה הווירטואלית לענייני הלכה, אחריהם סרטון flash מצחיק באנגלית על ספרנית-שוטרת (די ברור שתוצאה זו נכללה בגלל הטקסטים בקישורים), ספרייה מקוונת של אתר החדשות NFC (אגב, בדף עצמו כתוב "ספרייה" ולא "ספרייה"), והמלה "ספרייה" מופיעה שוב רק בקישורים) ודף הבית של ארגון יד שרה, שממנו יש קישור לספרייה הדיגיטאלית של הארגון. כנראה, שדף זה דורג גבוה כל כך מפני שיש קישורים רבים ברשת אל ארגון יד שרה, לאו דווקא בהקשר של הספרייה

טבלה מספר 1: שאילתות על ספרייה

מורפיקס	נענע	וואלה	Google	השאילתה	
952,26 מורפולוגי 3,541 מדויק	סיווג: ספריות 338 אתרים	99 אתרים מתוך המדריך המסווג 5,647 דפים	73,900	ספרייה	1
26,952 מורפולוגי 1,872 מדויק	סיווג: ספריות 101 אתרים	62 אתרים מתוך המדריך המסווג 2,231 דפים	40,500	ספרייה	2
26,952 מורפולוגי 2805 מדויק	31 אתרים	20 אתרים מתוך המדריך המסווג 4298 דפים	124,000	הספרייה	3
19,431 דפים בשני המקרים	13 אתרים	8 אתרים מתוך המדריך המסווג 585 דפים	7,530	ספרייה AND הספרייה	4
354,933 מורפולוגי 347,412 מדויק	אין אפשרות חיפוש כזאת	אין אפשרות חיפוש כזאת	171,000	ספרייה OR הספרייה	5
26,952 מורפולוגי 1,982 מדויק	35 אתרים	13 אתרים מתוך המדריך המסווג 585 דפים	65,600	הספרייה	6
347,412 מורפולוגי 1,925 מדויק	6 אתרים	4 אתרים מתוך המדריך המסווג 2,347 דפים	32,200	בספרייה	7
26,952 מורפולוגי 3,541 מדויק	6 אתרים (שונים מהאתרים בשאילתה 7)	6 אתרים מתוך המדריך המסווג 1,244 דפים	16,000	בספרייה	8
9,738 מורפולוגי 9385 מדויק	1,428 אתרים	4 אתרים מתוך המדריך המסווג 3081 דפים	51,600	ספרית	9
67 (בחיפוש ביטויים אין ניתוח מורפולוגי) 148	אתר אחד	7 דפים	26	"ספרייה לאומית"	10
	4 אתרים (אף אחד מהם אינו הספרייה הלאומית ואוניברסיטאית)	73 דפים	1,220	"הספרייה הלאומית"	11
0 מורפולוגי 0 מדויק	אין אפשרות חיפוש כזאת	אין אפשרות חיפוש כזאת	243,000	OR של כל השאילתות הקודמות: ספרייה OR ספרייה OR הספרייה OR הספרייה OR בספרייה OR בספרייה OR ספרית OR "ספרייה לאומית" OR "הספרייה הלאומית"	12

ולצורך הדירוג, מתחשב גוגל במספר הקישורים ובאיכות הקישורים המובילים לדף (Brin & Page, 1998). לא ציפינו ששתי תוצאות מאותו האתר יופיעו בין חמש התשובות הראשונות, לכן נחשיב רק אחת מהם כרלוונטית. גם הסרטון לא היה ממש רלוונטי בעינינו וגם לא דף הבית של יד שרה, כי ממנו יוצא קישור בודד אל ספרייה. precision@5 במקרה זה יהיה 40 אחוז.

שאלתה מספר 3 - "הספרייה":

התוצאה הראשונה היא דף הבית של הספרייה הלאומית, השנייה הספרייה הווירטואלית של מט"ח, השלישית והרביעית הן דפים מתוך הספרייה הישראלית למחול (בתוך בית אריאלה) והאחרונה היא דף הבית של הספרייה העירונית של קרית גת. רק אחת משתי התוצאות מאותו האתר היא בלתי רלוונטית, לכן precision@5 נקבעה כ-80 אחוז.

שאלתה מספר 4 - ספרייה AND הספרייה:

מציגה את הספרייה הווירטואלית של מט"ח, את הספרייה העירונית של קרית גת, ספריית תמונות בפורטל של חב"ד, הספרייה האזורית של מועצת מטה אשר ואת הספרייה של הפקולטה למשפטים של אוניברסיטת בר-אילן. שלוש מתוך חמש התוצאות אוחזרו גם בחלק מהשאלות הקודמות. במקרה זה, precision@5 היה 100 אחוז.

שאלתה מספר 5 - ספרייה OR הספרייה:

העלתה תוצאות שהוצגו כבר קודם: הספרייה הלאומית, הספרייה הווירטואלית של מט"ח, הספרייה הישראלית למחול, הספרייה הווירטואלית לענייני הלכה וספריית הפקולטה למשפטים של אוניברסיטת בר-אילן. במקרה זה, precision@5 היה 100 אחוז.

שאלתה מספר 6 - הספרייה:

שוב עולה הספרייה הווירטואלית של מט"ח, אתר של הספרייה של מכללת אשקלון, אחריו שני דפים מתוך אתר הספרייה של מכללת בית ברל ואתר הספרייה של מכון וינגייט. במקרה זה, precision@5 היה 80 אחוז.

שאלתה מספר 9 - ספרייה:

מעלה תוצאות שונות לגמרי. התוצאה הראשונה היא הוצאת ספריית הפועלים (כנראה שהדף חסר), השנייה הוא אחד מתתי הסיווגים בפורטל המורים הפרטיים בישראל (יש קישור ל"ספריית העצים" מהתפריט הצדדי), הספרייה של מכללת אורנים, ספריית מאמרים של נחשון - תכנון מערכות מזון (הפעם הקישור מהתפריט בתחתית הדף) וספריית ההשאלה של ספרי לימוד באוניברסיטת בן גוריון. מכיוון שהדפים מהם יוצא קישור בודד לספרייה אינם רלוונטיים, precision@5 היה 60 אחוז.

שאלתה מספר 12 - OR של כל השאלות הקודמות:

התוצאות הראשונות שעלו זהות כמעט לחלוטין לתוצאות הראשונות שהועלו במענה לשאלתה מספר 5, פרט לכך, שכאן מוצגים שני דפים מהספרייה הישראלית למחול (במקום הספרייה של הפקולטה למשפטים של אוניברסיטת בר-אילן); במקרה זה, precision@5 היה 80 אחוז.

לסיכום, השאלתה שדלתה את התוצאות הרלוונטיות ביותר בניסוח הקצר ביותר היא שאלתה מספר 5 (ספרייה OR הספרייה). ביחד עם זאת, בולטת היעדרותן של הספריות האקדמיות ושל הספריות הציבוריות הגדולות. במספר מקרים המלה המבוקשת הופיעה בתפריט צדדי או תחתית בלבד, כך שלדף עצמו לא היה שום קשר לנושא.

וואלה!

וואלה! מציג תחילה את האתרים מהמדריך המסווג (כאשר קיימים כאלה), נבדוק את האתרים מהמדריך.

שאלתה מספר 1:

התוצאות הראשונות הן: דף הבית של אוניברסיטת בר-אילן (לא של אחת הספריות), שבו מופיע קישור בודד לאתר הספריות, דף הבית של המכון הישראלי לייצוא (אין שום אזכור של ספרייה, אך בתיאור שניתן מופיעה המלה ספרייה), דף הבית של חנות הספרים הווירטואלית דיבוק (גם כאן אין אזכור של המלה ספרייה), ספרייה (כתיב מלא) פדגוגית קיבוץ סאסא (בתיאור מופיעה המלה ספרייה) ודף הבית של האוניברסיטה העברית (אך הכתובת לא נכונה או שאינה קיימת, ולכן לא ניתן להיכנס לאתר המצויין). לפי הכללים שנקבעו,

precision@5 היה 20 אחוז.

שאלתה מספר 2:

התוצאות הראשונות הן: התוצאה הראשונה מובילה ל- elibrary אתר באנגלית, אשר שמו תורגם על ידי העורכים ל"ספרייה אלקטרונית". תוצאה שנייה מפנה גם היא לאתר באנגלית: Jewish Virtual Library - לא רואים את המלה ספרייה בשום מקום (לא בדף ולא בתיאור), במקום השלישי מופיע שוב אתר בשפה האנגלית, אשר תורגם ל"ספרייה למפתחים ברשת". המקום הרביעי והחמישי מפנים את המחפש לאתר ספריית הקונגרס (אותו URL, אך שני תיאורים שונים), ובשניהם מופיעה המלה "ספריית" (ולא ספרייה). לפי התוצאות נראה שיש אולי התייחסות חלקית למורפולוגיה. אנחנו מעוניינים במידע בעברית, ועל כן precision@5 עומד במקרה הזה על 0 אחוז.

שאלתה מספר 3:

התוצאות הראשונות הן: התוצאה הראשונה היא האתר בשפה האנגלית, IPL (הספרייה הציבורית באינטרנט), אחריה הספרייה הוירטואלית של מט"ח, התוצאה השלישית והחמישית מפנות לספרייה הציבורית של ניו יורק (אחת לדף הבית של הספרייה והשנייה לאוסף הדיגיטאלי). התוצאה הרביעית מפנה אל "הספרייה הדיגיטאלית של קליפורניה". מעניין לציין, שהתוצאה הראשונה בחיפוש החופשי בוואלה! היא דף הבית של הספרייה הלאומית, אבל אתר זה אינו מוזכר כלל בתוצאות שאוחזרו מהמדריך המסווג. כמו במקרה הקודם, גם כאן precision@5 הוא 0 אחוז.

שאלתה מספר 4:

התוצאות הראשונות הן: הספרייה הציבורית באינטרנט, את הספרייה הציבורית של ניו יורק, דף הבית של בית צבי, בית הספר הגבוה לאמנויות הבמה (בתיאור מופיע "באתר תמצאו מידע כללי על הספרייה"), הספרייה העירונית של קרית גת, ודף הבית של הספרייה המרכזית לעיוורים בישראל. precision@5 הוא 40 אחוז.

שאלתה מספר 6:

התוצאות הראשונות מעלות את בית לוחמי הגטאות (הדף באנגלית, "הספרייה" מופיעה בתיאור), הספרייה הציבורית של ניו יורק, "הספרייה המדהימה" (אתר באנגלית), הספרייה המרכזית לעיוורים בישראל (בתיאור יש שימוש במלה "הספרייה" בעוד שבאתר כתובה המלה "הספריה"). התוצאה החמישית הייתה אמורה להוביל אל "הספרייה העירונית של נתניה", אך הדף אינו קיים. precision@5 הוא 40 אחוז.

שאלתה מספר 9:

התוצאות הראשונות מעלות את "ספריית בריאות ברשת", את חנות הספרים "דיבוק", מרפאה לחיות מחמד (יש בה מספרה לפי התיאור, אך האתר אינו קיים בפועל), ואתר של אורט סינגלובסקי ("קהילה וירטואלית בית ספרית"). למרות שהתוצאות לגיטימיות מבחינה לשונית, אנחנו רצינו מידע על libraries ולכן precision@5 הוא 0 אחוז.

נענע

במדריך המסווג של נענע קיים הסיווג "ספריות" (אוחזר כתוצאה של השאלתה "ספריה"). הסיווג מכיל 65 אתרים, הראשונים: אתר ה-ALA, אתר הספרייה הלאומית הצרפתית, הספרייה הדיגיטאלית של קליפורניה, ספרייה ציבורית בפייטסבורג והאוסף הדיגיטאלי בספרייה הציבורית בניו יורק; כלומר, כל האתרים הראשונים לועזיים מסודרים בסדר אלף-בית, האתר הראשון מישראל נמצא במקום השביעי, זהו אתר ספרייה ערבית לילדים בנצרת (האתר בערבית), במקום התשיעי מופיע אתר בית הספרים הלאומי.

שאלתה מספר 1:

חמשת האתרים הראשונים שהוצגו לשאלתה היו: דף שמפנה לרשימת ספריות (זהה לתוצאה הראשונה בגוגל), ספרייה ציבורית באפרת, אתר בתוך וואלה! שאינו קיים, אתר מרכז ההדרכה לספריות ציבוריות - libagent, שהוא יישום המאפשר להאריך השאלת ספרים בספריות אוניברסיטאיות, אלא שהוא לא קיים יותר. precision@5 הוא 60 אחוז.

שאלתה מספר 2:

העלתה רשימת אתרים שונה: התוצאה הראשונה הפעם היא הספרייה של הנשינול ג'אוגרפיק, אחריו מרכז ההדרכה לספריות ציבוריות בישראל, אתר ספריית הקונגרס, יישום libagent ואתר של ה-library corporation שהיא חברה מסחרית המספקת שירותים לספריות. precision@5 הוא 20 אחוז.

שאלתה מספר 3 - "ספרייה":

התוצאות הראשונות עלו כדלהלן: הספרייה הקולית של אוניברסיטת משיגן, דף הבית של הספרייה המרכזית של אוניברסיטת תל אביב, הספרייה הישראלית למחול (בבית אריאלה), הספרייה המרכזית לעיוורים והספרייה המרכזית של מועצה מקומית מטה אשר. precision@5 הוא 80 אחוז.

עבור שאלתה מספר 4 - "ספרייה AND הספרייה":

היו התוצאות הראשונות ספריית מט"ח, הספרייה הקולית של משיגן והספרייה הווירטואלית של אוניברסיטת דרום קליפורניה. שני האתרים באנגלית ובתיאורים בעברית מופיעה הצורה "הספרייה" בלבד - כך שלא ברור כיצד התפרשה השאלתה הזאת, רשימת הספריות (זהה לתוצאה הראשונה של גוגל עבור "ספרייה") והספרייה של מכללת שאנן. precision@5 הוא 60 אחוז.

עבור שאלתה מספר 6 - "ספרייה":

קיבלנו את הספרייה של הנשינול ג'אוגרפיק, אתר ברוסית - הספרייה של מקסים מושקו, הספרייה הלאומית הבריטית, הספרייה של מכללת שאנן והספרייה הווירטואלית של מט"ח. precision@5 הוא 40 אחוז.

שאלתה מספר 9 - "ספרייה":

אחזרה את האתרים הבאים: ספריית גופנים, אתר של הוצאת הספרים "הקיבוץ המאוחד", מספרת ילדים בשם "קצוץ", מספרה לכלבים וספריית DVD. המלה "ספרית" יכולה להתפרש כ"מעצבת שיער" - כנראה שזו הסיבה שקיבלנו גם מספרות. precision@5 הוא 20 אחוז.

מהבדיקה עולה, שייתכן שהאתרים וואלה! ונענע מבצעים איחוד צורות חלקי, מה שיכול להסביר את ההבדלים בין התוצאות לשאלות השונות, וגם את המקרים בהם מלת החיפוש לא הופיעה בצורתה המדויקת לא בדף ולא בתיאור של עורכי המדריכים. השאלתה המעניינת ביותר מבחינת פירושים היא "ספרית" אבל, כאמור, ייתכן שמדובר במעצבת שיער או בשם תואר (כמו בביטוי פעילות בית ספרית).

מורפיקס

החיפוש המורפולוגי העלה תוצאות זהות עבור השאלות: "ספרייה", "ספרייה", "הספרייה", "הספרייה", "ובספרייה". התוצאה הראשונה בכלל לא קיימת באינטרנט, השנייה היא דף הבית של הספרייה הרפואית באוניברסיטת בן גוריון ושל המרכז הרפואי סורוקה, התוצאה השלישית היא אתר של נגרייה (הבונה ספריות), התוצאה הבאה היא דף הבית של מרכז ההדרכה לספריות בישראל והתוצאה החמישית היא שוב אתר שאינו קיים כלל. precision@5 הוא 40 אחוז.

החיפוש המדויק (ללא צורות מורפולוגיות) עבור "ספרייה" העלה שני דפים ראשונים שאינם קיימים, תרגיל ספרייה בפקולטה לחקלאות באוניברסיטה העברית (תוצאה זו מופיעה פעמיים תחת שתי כתובות שונות) ודף הסבר על תקליטור בשם "ספרייה תורנית ממוחשבת 2". precision@5 הוא 20 אחוז.

שאלתה מספר 2 - "ספרייה":

חיפוש מדויק. התוצאות הראשונות שקיבלנו במקרה זה היו: שני דפים שאינם קיימים, הספרייה למוזיקה ממכללת ליונסקי, דף שכתוב עליו "ספרייה", ופורום "קוראים מגיבים" של תיכון מקרית חיים. precision@5 הוא 20 אחוז.

שאלתה מספר 3 - "ספרייה":

חיפוש מדויק. העלתה תוצאות הכוללות את הספרייה הרפואית של אוניברסיטת בן גוריון (תוצאה זו נכללה בין חמשת התוצאות הראשונות גם עבור החיפוש המורפולוגי), הספרייה המרכזית לעיוורים, הספרייה האזורית של מועצת מטה אשר, ושתי הפניות (URL זהה) לספרייה לתקשורת של המכללה למנהל (הדף אינו קיים יותר). precision@5 הוא 60 אחוז.

שאלתה מספר 4 ושאלתה מספר 5:

הן שאלות בוליאניות. התוצאות שקיבלנו הן מוזרות. במקרה של שאלת ה-AND לא היה הבדל בין החיפוש

המורפולוגי לחיפוש המדויק. עבור החיפוש המדויק קיבלנו מספר בלתי סביר של תוצאות (מספר התוצאות אמור להיות קטן יותר ממספר התוצאות עבור כל אחד מהחיפושים המדויקים הנפרדים). בשאלת OR קיבלנו בשני המקרים מספר תוצאות שהיה גדול בהרבה מהמצופה (לכל היותר סכום התוצאות הנפרדות). הסיבה לתוצאות המוזרות האלה היא כנראה שבמקרים אלה מתבצע ניתוח מורפולוגי רחב יותר. אנו מבססים הנחה זו על כך שבדף התשובות מודגשת הפעם גם המלה "ספר" ולא רק צורות שונות של "ספרייה". עבור שאילתת ה-AND, התקבלו בין חמשת התוצאות הראשונות שני דפים שאינם קיימים (אחד היה של ספרייה והשני של נגרייה) ושלושה דפים של בתי ספר (בית ספר אחד מופיע פעמיים עם URL זהה). עבור התוצאות המדויקות של שאילתת ה-OR, קיבלנו את בית הספר למנהל עסקים של האוניברסיטה העברית ועוד ארבעה דפים שאינם קיימים (של בתי ספר). עבור החיפוש המורפולוגי, הועלו שני דפים שאינם קיימים, דף הספרייה הרפואית של אוניברסיטת בן גוריון, מרכז ההדרכה לספריות ואתר של נגרייה (התוצאות הראשונות זהות בדיוק לחיפוש המורפולוגי עבור וריאציות שונות של "ספרייה", אשר תוארו בתחילת הסעיף). גם תוצאות שאילתת מספר 12 מצביעות על כך שמורפיקס אינו מתמודד נכון עם אופרטורים בוליאניים. הדיוק של חמש התוצאות הראשונות היה 0 אחוז בכל המקרים פרט ל-OR המדויק: שם, precision@5 היה 40 אחוז.

שאלתה מספר 6 - חיפוש מדויק של "הספרייה":

חמשת התוצאות הראשונות שהועלו מכילות שלושה דפים שאינם קיימים (אחת מהן הפנייה לא נכונה לספרייה הווירטואלית של מט"ח), ושניים נוספים עם URL זהה המפנים כנראה לתרגיל במחשב, הפניה לביצוע תרגיל במחשב - דף זה אמנם קיים, אך אינו מכיל את המלה מבוקשת, ודף שמסביר על הספר העברי המקוון במסגרת ספרייה וירטואלית. precision@5 הוא 20 אחוז.

לא ברור כלל מדוע יש הבדל בחיפוש המורפולוגי בין "בספרייה" לבין "בספרייה". במקרה זה, אפילו התוצאות הראשונות אינן דומות. לפי ההדגשות בתוך התוצאות מתברר ש"בספרייה" מורחב ל"ספר" בעוד ש"בספרייה" מביא רק וריאציות של "ספרייה".

שאלתה מספר 9 - "ספרייה":

חיפוש מורפולוגי. הפעם מעלה החיפוש דף של אתר ספריית קצרין, ועוד ארבעה דפים שאינם קיימים (שלושה מתוכם מתייחסים לפעילות בית ספרית, והרביעי היה אמור להוביל לדף בספרייה של מכללת תל חי). בחיפוש המדויק עולות ארבע תוצאות ראשונות שהן זהות לארבע התוצאות הראשונות של החיפוש המורפולוגי, והתוצאה החמישית, במקום הספרייה מכללת תל חי אמורה להוביל ללוח הודעות של האוניברסיטה העברית, אך הכניסה דורשת סיסמה. precision@5 הוא 20 אחוז.

תוצאות החיפוש במורפיקס מאכזבות מאוד. מכיון שנוכחנו כבר קודם שהמאגר אינו מעודכן, ראוי היה שמפעילי יוסיפו הודעה שהמאגר הקיים הוא לצורך הדגמת היכולות המורפולוגיות בלבד ואינו מתעדכן באופו שוטף. ביחד עם זאת ראינו גם מספר מוזרויות הקשורות בטיפול המורפולוגי. כמו כן, האופרטורים הבוליאניים אינם פועלים כראוי.

עבור כל אחת מהשאלות שבדקנו באופן מעמיק, ציינו גם את הערך של precision@5. ערך זה אינו מספיק כדי להעריך את התוצאות כשלעצמו, כי הוא אינו מרמז על איכות התוצאות כגון הופעה או אי-הופעה של דפי הבית של ספריות אקדמיות מרכזיות ושל ספריות ציבוריות גדולות בין חמש התוצאות הראשונות.

שאלות נוספות

השימוש בגרשיים בשפה העברית גורם לקושי נוסף, מכיון שהסימן גרשיים משמש ברוב כלים לציון חיפוש ביטוי. דוגמאות לתוצאות מוזרות המתקבלות כתוצאה מכך ניתן לראות בהרצת השאלות "נוה אטי"ב" (לגרשיים האמצעיים אמורה להיות משמעות שונה מאשר לגרשיים בתחילת הביטוי ובסימו). במקרה זה גוגל מתבלבל לחלוטין כאשר מחפשים את הביטוי המדויק (ס תוצאות, למרות שקיימות תוצאות מתאימות במאגר) וואלה! מחפש את הביטוי ללא הגרשיים (נוה אטיב), נענע מוצאת תוצאות רק כאשר כותבים את הביטוי ללא גרשיים ומורפיקס מטפל נכון במקרה זה. שאילתה נוספת מעניינת היא "ניו" (שכונה בירושלים) - שאילתה המחזירה תוצאות המכילות את הצירוף נ', ו', ת', כגון יחצ"ניות וכו' בגוגל בוואלה ובנענע (פה מקבלים גם "מדיניות" וגם "תוכניות"). החיפוש המדויק של מורפיקס מחזיר תוצאות טובות (מבחינה לשונית) גם במקרה זה.

סיכום ומסקנות

כמות המידע בשפה העברית ב- Web הולכת וגדלה כל הזמן. ללא כלי אחזור מתאימים לא נוכל לנצל את המידע הרב הטמון בה, נטבע בים המידע הזה ונחוש שאנו הולכים לאיבוד בין קשרי ההיפרטקסט המעברים אותנו מדף אחד למשנהו.

עבור השפה האנגלית קיימים כלי חיפוש מתקדמים, בעיקר משום שהמאמץ המחקרי מתמקד כיום בשפה האנגלית. אבל, כלים שפותחו עבור השפה האנגלית מתאימים רק באופן חלקי עבור משתמשי הרשת בעברית. הסיבה המרכזית לכך היא שאנגלית היא שפה פשוטה מבחינה מורפולוגית בעוד שעברית היא שפה מורכבת ביותר. הגורמים המקשים העיקריים הם צירוף התחיליות, השמטת התנועות (כתוצאה מכך ריבוי משמעויות ומספר צורות כתיבה נכונות - כתיב מלא וכתיב חסר). מבחינה מחקרית קיימים פתרונות (כפי שראינו בסקירת הספרות), אך שיטות אלה אינן באות לביטוי לרוב בכלים המסחריים, כנראה בעיקר מסיבות כלכליות. הכלי החופשי היחיד כיום אשר מתמודד עם בהצלחה חלקית עם האתגרים של השפה העברית (מורפיקס) אינו מתמודד עם האתגרים האחרים של ה- Web: התחדשות מתמדת, השמטת כפילות מידע והצורך בדירוג טוב של התוצאות. הגוגל, שהוא כלי החיפוש הפופולארי ביותר בקרב גולשי האינטרנט בישראל, מספק את התוצאות הרחבות והעדכניות ביותר. בשלב זה נראה ששילוב של גוגל עם מורפיקס משופר (פתרון הבעיות המורפולוגיות שעליהן הצבענו בגוף המאמר) היה משפר מאוד את חווית הגלישה של המשתמש המעוניין במידע בשפה העברית.

לסיום נתייחס גם לתוצאות הראשונות של חיפוש המידע עבור מונחים שונים הקשורים בספריות. הספרייה הלאומית והאוניברסיטאית (או בשמה הרשמי "בית הספרים הלאומי והאוניברסיטאי") אינו ממופתחת כלל במדריך של וואלה! וגם לא במורפיקס (לפחות לא הצלחנו למצוא הפניה ישירה בין כמה עשרות התוצאות הראשונות), וניתן למצוא אותה בנענע רק כאשר מחפשים את "בית הספרים הלאומי" או כאשר מדפדים בסיווג "ספריות" עד למקום התשיעי. מעט ספריות אוניברסיטאיות הופיעו בין התוצאות הראשונות (הספרייה של אוניברסיטת תל אביב, הספרייה של אוניברסיטת בן גוריון והספרייה של אוניברסיטת בר-אילן) ופרט להפניה לספרייה של אוניברסיטת תל אביב, לא הופנה המחפש לספריות המרכזיות או לדפים המרכזיים של אתרי הספריות הללו. דפים מאוניברסיטת חיפה הופיעו עבור השאלית "בספרייה" במורפיקס-דפים אלו הובילו לאתר הישן של הספרייה של אוניברסיטת חיפה. הוזכרו מספר ספריות של מכללות: אורנים, אשקלון, בית ברל, המכון הטכנולוגי חולון, וינגייט, המכללה למנהל (ההפניה הובילה לדף שאינו קיים), מכללת לוינסקי (הספרייה למוזיקה), מכללת שאנן ומכללת תל חי (קישור לדף שאינו קיים). הספריות הציבוריות שהועלו בחיפושם הן: בית אריאלה הוזכרה באמצעות הספרייה למחול בלבד, הספרייה המרכזית לעיוורים, הספרייה העירונית באפרת, בנתניה (הקישור לא פעל) ובקרית גת והספרייה האזורית של מועצת מטה אשר. מרכז הדרכה לספריות הופיע בין התוצאות הראשונות בנענע ובמורפיקס, אך אינו ממופתח במדריך המסווג של וואלה!. באופן כללי הורגש שעורכי המדריכים המסווגים (וואלה! ונענע) מעדיפים למפתח אתרים לועזיים. המלצתנו היא שהספריות המרכזיות יפעלו כדי להבליט את נוכחותן במנועי החיפוש ובמדריכים המרכזיים. נוף מנועי החיפוש ב- Web עובר שינויים תמידיים (כלים מתווספים, נעלמים או משנים את היכולות שלהם), וגם מידת הפופולאריות של הכלים משתנה עם הזמן וכו'). לדוגמה, ב- 1 בפברואר 2005, Microsoft חנכה את מנוע החיפוש החדש שלה (<http://search.msn.com>), לרגל אירוע זה נערכה ב- ynet (2005) השוואה בין מנועי חיפוש אשר מסוגלים לחפש בשפה העברית (google, msn, yahoo, וואלה! ומורפיקס). לא נבחנו היכולות מבחינת השפה העברית, אלא מספר התוצאות והרלוונטיות שלהן (אין פירוט) על ביטויים ושמות בחדשות.

לאור השינויים התכופים, מומלץ על עריכת בדיקות תקופתיות לצורך בחינת יכולות הכלים באחזור מידע בשפה העברית. בנוסף על זאת, רצוי לערוך מחקרי שימוש ולבחון כיצד מחפשים משתמשי ה- Web בשפה העברית ומתן התכונות החשובות להם במיוחד בכלים השונים.

מקורות

- האקדמיה ללשון העברית (2004). *מילון-מידע* פרק 35: רישות (תשס"ד). אוחר ב-28, ינואר, 2005
<http://hebrew-terms.huji.ac.il/milonimsearch3.asp?milonid=232&ord=ktaim.ketae> מ
- TNS/טלסקר (2004). *החשיפה לאתרי האינטרנט ע"פ סקר TIM – סתיו 2004*. אוחר ב 30, ינואר, 2005
<http://www.nrg.co.il/images/stuff/computers/sekertim.doc> מ
- מור, גל (12 בדצמבר, 2004). סקר: 3.2 מיליון גולשים בישראל. *ynet*. אוחר ב-30, ינואר, 2005 מ
<http://www.ynet.co.il/articles/0,7340,L-3017304,00.html>
- מור, גל (2 בפברואר, 2005). בדיקה השוואתית: MSN נגד גוגל. *Ynet*. אוחר ב-3, פברואר, 2005 מ
<http://www.ynet.co.il/articles/0,7340,L-3040604,00.html>
- מורפיקס (2001). *על מורפיקס*. אוחר ב-1, פברואר, 2005 מ <http://www.morfix.co.il/al-morfix.htm>
- מלינגו (2003). *מורפיקס. מנוע החיפוש האולטימטיבי עבור שפות מורפולוגיות*. אוחר ב-30, ינואר, 2005
http://www.melingo.co.il/hb_morfix_ab.htm מ
- מרגלית, א. (2004). השואה בין גישות שונות למימוש ישומי אחזור תוך שימוש במורפולוגיה עברית. *עלון SIGTRS* 10(2). אוחר ב-30, ינואר, 2005 מ <http://sigtrs.huji.ac.il/102-heb-morphology.pdf>
- רוזן, י. (1996). למה כל כך קשה לכתוב בעברית? *מעשה חושב*, יולי 1996. אוחר ב-30, ינואר, 2005
<http://sigtrs.huji.ac.il/qsm/qashetab.htm> מ
- Attar, A., Choueka, Y., Dershowitz, N., & Fraenkel, A. S. (1978). KEDMA - Linguistic tools for retrieval systems. *Journal of the Association for Computing Machinery*, 25(1), 52-66.
- Bar-Ilan, J., & Gutman, T. (2005). How do search engines respond to some non-English queries? *Journal of Information Science* 31(1), (2005) pp. 13-28.
- Berners-Lee, T. (1992). *W3 servers*. Retrieved January 29, 2005, from http://geonic.net/index.php?section=history&subsection=www&exe=w3_serv
- Brin, S., & Page, L. (1998). The anatomy of a large-scale hypertextual Web search engine. In *Proceedings of the 7th International World Wide Web Conference*, April 1998. Retrieved January 15, 2005, from <http://www-db.stanford.edu/pub/papers/google.pdf>
- Choueka, Y. (2005). The complexity of processing natural languages by computers. *SIGTRIS* 11(2) Retrieved January 30, 2005, from <http://sigtrs.huji.ac.il/112-ibud/112-ibud.files/frame.htm>
- Dapey Reshet (1999). *Hebrew FAQ*. Retrieved January 29, 2005, from <http://dapey.reshet.co.il/help/2067.htm>
- Fallows, D. (2005). *Search engine users*. Retrieved January 30, 2005, from http://www.pewinternet.org/pdfs/PIP_Searchengine_users.pdf
- Global Reach (2004). *Global Internet statistics (by language)*. Retrieved January 30, 2005, from <http://www.global-reach.biz/globstats/index.php3>
- Guggenheim, E., & Bar-Ilan, J. (to appear). Tauglichkeit von Suchmaschinen für deutschsprachige Abfragen. *Information, Wissenschaft und Praxis*, 56(1), 35-40.

iGuide (no date). *Israeli Internet Guide: Sites sorted by name*. Retrieved January 29, 2005, from <http://www.iguide.co.il/sites/sites.htm>

Moukdad, H. (2004). Lost in Cyberspace: How do search engines handle Arabic queries? In *Access to Information: Technologies, Skills, and Socio-Political Context. Proceedings of the 32nd Annual Conference of the Canadian Association for Information Science*, Winnipeg, June 3-5, 2004. Retrieved January 29, 2005, from http://www.cais-acsi.ca/proceedings/2004/moukdad_2004.pdf

Silverstein, C., Henzinger, M., Marais, H., & Moricz, M. (1999). Analysis of a very large Web search engine query log. *ACM SIGIR Forum*, 33(1). Retrieved February 1, 2005, from <http://www.acm.org/sigir/forum/F99/Silverstein.pdf>

Spink, A., Wolfram, D., Jansen, M. B. J., & Saracevic, T. (2001). Searching the Web: The public and their queries. *Journal of the American Society for Information Science and Technology*, 52(3), 226-234.

Sullivan, D. (2004). *comScore Media Metrix search engine ratings*. Retrieved January 30, 2005, from <http://searchenginewatch.com/reports/article.php/2156431>

Wintner, S. (2004). Hebrew computational linguistics: *Past and future*. *Artificial Intelligence Review*, 21(2), 113-138. Retrieved January 30, 2005, from <http://cs.haifa.ac.il/~shuly/publications/hcl.pdf>

