בלשנות חישובית עברית: עבר ועתיד

שולי וינטנר

1 בלשנות חישובית ועיבוד שפות טבעיות

בלשנות חישובית (computational linguistics) היא תחום מחקר המקשר בין בלשנות ובין מדעי המחשב. ישנן שתי דרכים שונות להתבונן בבלשנות חישובית: 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 תוכניות 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.

מאמר זה מתמקד בעיבוד שפות טבעיות, כלומר ביישומים של מדעי המחשב המצריכים ידע לשוני, או מדמים פעולות לשוניות. ישנן דוגמות רבות ליישומים כאלה, והן כוללות תרגום אוטומטי משפה אחת לאחרת; המרה של דיבור לכתב ולהיפך; מימשקים בשפה טבעית למערכות ממוחשבות; תמצות אוטומטי של מסמכים; בדיקת איות וסגנון; ועוד כהנה וכהנה. החשיבות של התחום היא עצומה: תקשורת עם מערכות ממוחשבות בשפה טבעית תהפוך מערכות כאלו לנגישות יותר, ותאפשר ליותר משתמשים, כולל משתמשים שאינם מומחים, ליהנות מפירות המערכת הממוחשבת; הבנה ויצירה של דיבור יניבו מערכות למענה קולי אוטומטי ויאפשרו הפעלה של מכשירים ללא מגע; המרת דיבור לכתב תאפשר לחירשים "לשמוע" שיחות טלפון; המרת כתב לדיבור תאפשר לעיוורים "לקרוא" את הדואר האלקטרוני שלהם; תרגום אוטומטי יחסוך סכומי

עתק אם ניתן יהיה לכתוב עותק אחד של מדריך השימוש במכשיר כלשהו, ולהפיק באופן אוטומטי תרגומים לשפות כל המדינות בהן נמכר המכשיר. האפשרויות בתחום אינן מוגבלות, אך מצב המחקר הנוכחי הוא כזה שרק מעט מהיישומים האפשריים אכן קיימים, ואיכות הקיימים אינה מספקת.

2 האתגר

כדי לבחון את מצב המחקר הנוכחי בעיבוד שפות טבעיות, כדאי לבחור בשפה שזכתה להתעניינות מְרַבית: השפה האנגלית. בעידן האינטרנט קל למצוא יישומים המשלבים בתוכם טכניקות של עיבוד שפות טבעיות, שכל החפץ בכך יכול להפעילם. שני יישומים נפוצים במיוחד ושימושיים במיוחד הם הבנת שאלות ותרגום אוטומטי.

מנועי חיפוש באינטרנט הם יישומי מחשב המאפשרים למשתמש לפרט מספר מילות מפתח, ולעתים צירופים שלמים, ומספקים הפניות לאתרים שבהם מסמכים המכילים את מילות המפתח הללו. התועלת שמנועי חיפוש יכולים להפיק מטכנולוגיה של עיבוד שפות טבעיות היא ברורה: במקרים רבים ניתן למצוא אלפי, ואף רבבות, מסמכים העונים לדרישות המשתמש, וכדי לברור את הנכונים והרלוונטיים מביניהם, על המערכת "להבין" את כוונת המשתמש באופן מדויק. לדוגמה, אם ברצוני לחפש מידע על ספרי תורה, אפשר שאספק למנוע החיפוש שתי מילות מפתח: ספרים + תורה. במקרה כזה, איני מעוניין לקבל כתשובה הפניה לאתר שבו מידע על תורת המספרים. כדי להתמודד עם הבעיה, קיימים אתרים באינטרנט בהם ניתן לשאול את המערכת שאלה בשפה טבעית, ולא רק לספק לה מילות מפתח. אתר אחד כזה הוא http://www.ask.com, והדוגמות הבאות נלקחו ממנו. האתר מספק הפניות למסמכים בהם טמונה התשובה, וההפניות מוצגות כשאלות דומות לשאלה המקורית; כל הפניה היא קישור למסמך שבו נמצאת התשובה לשאלה.

כשמפנים למנוע החיפוש את השאלה הבאה:

Who was the second president of the United States?

התשובות המתקבלות, כצפוי, הן נכונות ומדויקות (כלומר, ההפניות שמספק המנוע עונות על שאלות דומות מאוד לשאלה המקורית):

Where can I find information about U.S. President John Adams?
Where can I find encyclopedic biographical resources on John Adams?

ועוד הפניות דומות. לעומת זאת, כשמפנים למנוע החיפוש את השאלה:

Who was the United States president following Washington?

מתקבלות ההפניות הבאות:

Where can I learn about the historical political figure George Washington?
Where can I find information about U.S. President George Washington?

במילים אחרות, נראה שהתוכנה אותה מפעיל מנוע החיפוש אינה "מבינה" את משמעות השאלה; במקום להבין, נראה שהתוכנה מסתמכת על התאמה במילות מפתח, ומכיוון שהמונח "George Washington" מופיע בשאלה, הוא מופיע גם בתשובות שהמערכת מספקת.

ניסוי אחר ניתן לערוך על ידי שימוש בתוכנת תרגום אוטומטי של חברת Systran, הפתוחה לשימוש באתר http://babelfish.altavista.com/. התוכנית מאפשרת תרגום של מסמכים מאנגלית לשמונה שפות אחרות, ומהן לאנגלית. כדי לבחון את ביצועי התוכנה, השתמשנו בה כדי לתרגם את הפסקה הראשונה מתוך אליס בארץ הפלאות ללואיס קרול מאנגלית לאיטלקית. המקור הוא:

Alice was beginning to get very tired of sitting by her sister on the bank, and of having nothing to do: once or twice she had peeped into the book her sister was reading, but it had no pictures or conversations in it, 'and what is the use of a book,' thought Alice 'without pictures or conversation?'

התרגום לאיטלקית שהציעה התוכנה הוא:

Alice stava cominciando ad ottenere molto tired di seduta dalla sua sorella sulla banca e di non avere niente fare: una volta che o avesse pigolato due volte nel libro che la sua sorella stava leggendo, ma non avesse immagini o conversazioni in esso, 'e che cosa l'uso d'un libro,' pensiero Alice 'senza immagini o conversazione?'

כדי לבחון את איכות התרגום, תרגמנו את הפסקה שלעיל האיטלקית לאנגלית באמצעות אותה תוכנה. התוצאה שהתקבלה היא:

> Alice was beginning to very obtain tired of sitting from its sister on the bank and not to have nothing to make: once that or he had pigolato two times in the book that its sister was reading, but did not have images or conversations in it, 'and that what is the use of a book,' thought Alice 'without images or conversation?'

מהי הסיבה לאיכות הירודה של ביצועי שתי המערכות הללו, הן הבנת השאלות והן התרגום האוטומטי? מהם הקשיים הטמונים ב"הבנה" של שאלות בשפה טבעית, או בתרגום משפה אחת לאחרת? התשובה לשאלות אלו היא ששפות טבעיות הן קשות מטבען. התהליכים הקוגניטיביים הכרוכים בהבנה וביצירה של מבעים בשפה טבעית הם כנראה מורכבים מאין כמותם, ובכל אופן אין בידינו כלים להבינם כיום. כתוצאה מכך, הניסיונות לחקות באמצעות מחשב חלק מהתהליכים הללו נתקלים בקשיים מובנים. כדי להמחיש את העומק והמורכבות של שפה טבעית, נסקור להלן מספר שלבים נחוצים בדרך למימוש יישום מורכב, למשל תרגום אוטומטי, באמצעות מחשב. הסקירה שאינה תלוית שפה מאורגנת לפי רבדים שונים של עיבוד לשוני, המקבילים לרבדים שונים של מידע בלשני, והיא מתמקדת בתופעה אחת המקשה על עיבוד ממוחשב בכל אחד מהרבדים: בעיית ריבוי המשמעות. להלן נשתמש במונח ריבוי משמעות (ambiguity) כדי לציין תופעות בהן צורה אחת (למשל, מילה או משפט) ניתנת לניתוח במספר אופנים שונים, כך שניתן לייחס לה יותר מייצוג אחד.

כדאי להעיר כאן כי הסקירה שלהלן מתייחסת ליישומים מורכבים; בשנים האחרונות מופעלת טכנולוגיה של עיבוד שפות טבעיות בהצלחה מרובה ביישומים המצריכים הבנה של שפה ברמה נמוכה יותר. דוגמות ליישומים כאלה כוללות סיווג של מסמכים לקטגוריות על־פי נושאיהם; תמצות אוטומטי של מסמכים (summarization); הפניה אוטומטית של דואר אלקטרוני שאדם הצריך לטפל בו על־פי תוכן המסר; ועוד כהנה וכהנה.

תורת ההגה

תורת ההגה היא תחום המחקר הבלשני העוסק בהיגוי, הן בפונטיקה העוסקת בצלילים המופקים על ידי איברי הדיבור שלנו כשאנו מדברים, תכונותיהם הפיסיות, תפיסתם באיברי השמיעה והן בפונולוגיה, החוקרת את הצלילים האלה ברמה מופשטת, מגדירה אותם ואת צירופיהם ובוחנת את תפקודם במערכת הלשונית. מערכת ממוחשבת שהקלט או הפלט שלה הם שפה מדוברת חייבת לכלול ידע פונולוגי כדי לטפל נכונה בהגאים. ריבוי משמעות ברמה הפונולוגית

מתבטא באופנים שונים: הדוגמה הנפוצה היא הומופונים, או מילים שונות הנהגות באופן זהה, כמו למשל המילים week ו־weak באנגלית או כלה וקלה בעברית. בעברית הבעיה מעניינת עוד יותר, שכן לפעמים מילה אחת נהגית באופן זהה לצירוף המורכב מיותר ממילה אחת, כגון הקלה (שם פעולה של הפועל הקל) לעומת הקלה (צורת נקבה מיודעת של התואר קל) או שלו (כינוי הקניין של בנטיית גוף שלישי יחיד) לעומת שלא (מילת השימוש ש לפני מילת השלילה לא).

בעיות אחרות הקשורות בתורת ההגה כוללות אלופונים (הגאים זהים המתבצעים באופן שונה כתלות בסביבתם, כמו למשל הפונמות /p,k,b/ הנהגות f,x,v בהופיען אחרי תנועה), הגאים המתבצעים באופן שונה כתלות בדובר (למשל, ההבדלים בביצוע העיצורים הגרוניים בין דוברים ישראלים שונים) ועוד.

מורפולוגיה

המורפולוגיה חוקרת את מבנה המילים. כמעט לא ניתן להעלות על הדעת יישום ממוחשב של עיבוד שפות טבעיות שלא יצריך ידע מורפולוגי: כמעט כל יישום יצריך לכל הפחות מילון, ומערכת המיועדת לבצע חיפוש אינטליגנטי באינטרנט, למשל, תזדקק גם למנתח מורפולוגי (morphological analyzer) על מנת למצות את צורת הבסיס מתוך צורות נטויות של מילים, המופיעות במסמכים ברשת. כאן בולט במיוחד ריבוי המשמעות, והוא נובע הן מתהליכי גזירה (derivation) והן מתהליכי נטייה (inflection). למשל, צורן הסיום -י בעברית משמש בשני אופנים: הן לציון שייכות, גוף ראשון יחיד, והן להפיכת שם עצם לשם תואר. לפיכך, למילה ביתי שתי משמעויות: הבית שלי או הרגשה של בית. אחת הדוגמות המוצלחות ביותר לריבוי משמעות מורפולוגי בעברית היא התבנית שמנה, שלה לפחות תריסר משמעויות שונות (למשל, צורת נקבה של שם התואר שמן, או שם העצם שמן בצירוף כינוי קניין חבור, או מילת השעבוד ש־ ואחריה הפועל מנה, במשמעות ספר, או אף מילת השעבוד ש־, אחריה שם העצם מן שלו מצורף כינוי קניין חבור).

תחביר

אחד מתחומי המחקר העיקריים בבלשנות הוא התחביר (syntax), העוסק בשאלות של הצטרפות המילים לצירופים והצטרפות הצירופים למשפטים. התחביר מייחס מבנה לכל מבע בשפה, וגם הוא מעמיד שאלות כבדות של

ריבוי משמעות. לדוגמה, במשפט קיבלתי עניבה מאשתי משמש צירוף היחס מאשתי כתיאור של הפועל קיבלתי, בעוד שבמשפט קיבלתי עניבה ממשי משמש צירוף היחס ממשי כלוואי של שם העצם עניבה. המבנה התחבירי של שני משפטים דומים אלה סביר שיהיה שונה, שכן ההבדלים בין המשפטים צריכים יהיו, קרוב לוודאי, להשתקף במשמעות. לפיכך כל ניסיון של מנתח תחבירי לנתח את המשפט קיבלתי עניבה מאיטליה יידרש להתמודד עם שני המבנים האפשריים של המשפט: מאיטליה יכול להתפרש כלוואי של עניבה או כתיאור של קיבלתי.

תופעות רבות בתחביר של שפות טבעיות גורמות לקשיים בעיבוד ממוחשב. נציין כאן עוד שתיים בלבד: צירופי איחוי גורמים לעתים קרובות לריבוי משמעות. למשל, במשפט מאמר זה מתאר מחקר הקשור להבנה ויצירה של שפה טבעית, עשויה מילת החיבור ו להתפרש נכון, כמחברת את שמות הפעולה הבנה ויצירה; אך היא יכולה להתפרש גם באופן שגוי, כמחברת את שם הפעולה הבנה עם הצירוף יצירה של שפה טבעית. כמובן, ההבדל בין השניים עשוי להתבטא בתרגום לא נכון לשפה אחרת, או ביצירת משמעות שגויה. תופעה אחרת קשורה לצירופים פועליים משועבדים, כגון אלו שעשויים להשלים את המשפט העותר ביקש מבית המשפט... אם ההמשך הוא הצירוף הפועלי להשתחרר בערבות, הרי שהנושא המובלע של צירוף זה הוא העותר; אולם אם ההמשך הוא לפסוק לטובתו, הנושא המשתמע של הצירוף הפועלי המשועבד הוא בית המשפט. גם כאן, על מערכת ממוחשבת לעיבוד שפות להכריע בין המבנים השונים.

סמנטיקה

הסמנטיקה עוסקת במשמעות של מילים, צירופים ומשפטים בשפה. ריבוי המשמעות מתחיל כבר במילון: רבות הן המילים שלהן מספר משמעויות שונות, לעתים נבדלות לגמרי (כגון המילה חבל, במשמעות של חוט עבה או אזור גיאוגרפי), ולעתים נבדלות בדקות, כגון המילה פגישה שעשויה להיתרגם לאנגלית כ־meeting, appointment, date, או rendez-vous, בהתאם להקשר. אך מעניין עוד יותר ריבוי המשמעות העמוק, למשל זה הנעוץ באפשרויות שונות לחישוב הטווח של כמתים בשפה. למשל, השאלה מי מחברי הדירקטוריון המליץ על כל אחד מהמועמדים? עשויה לקבל שתי תשובות שונות: שם אחד, של אחד מחברי הדירקטוריון, אשר המליץ על כל המועמדים; או פירוט של כל חברי הדירקטוריון וליד כל אחד, שמות המועמדים עליהם המליץ.

בעיות אחרות בעיבוד סמנטי קשורות בפתרון של התייחסויות פרונומינליות, או אנאפורות. למשל, במשפט התחביר מייחס מבנה לכל מבע בשפה, וגם הוא מעמיד שאלות כבדות של ריבוי משמעות, רומז הכינוי הוא אל צירוף שמני, זכר יחיד, שהופיע קודם לכן במשפט. הצירוף הוא התחביר, אך אין מניעה ברורה לפרש את ההתייחסות כאילו היא רומזת אל הצירוף מבנה או אף אל הצירוף מבע. כמובן, מערכת ממוחשבת שלא תדע לפתור אנאפורות מסוג זה תיקלע לקשיים חמורים.

פרגמטיקה

גם כאשר למשפט משמעות ברורה, לעתים יש לו משמעויות משניות, התלויות בהקשר שבו המשפט מבוטא; הפרגמטיקה חוקרת תופעות אלו. למשל, המשפט אני אפגוש אותך מחר עשוי להתפרש כמשפט חיווי, המציין עובדה; אך בהקשר המתאים, הוא עשוי להתפרש כהבטחה; ובהקשר אחר הוא עשוי אף להוות איום. תופעה מעניינת נוספת כרוכה בהנחות יסוד (presupposition). מהמשפט מלך צרפת הוא קירח משתמעת ההנחה שלצרפת יש מלך; מהמשפט אני מתחרט שהחלטתי לנסוע לצרפת משתמעת ההנחה שהחלטתי לנסוע לצרפת. מערכת ממוחשבת שאמורה להסיק מסקנות על סמך קלט בשפה טבעית תיאלץ לחשב הנחות יסוד אלו כדי להגיע למסקנות הנכונות. תופעות אחרות שעוסקת בהן הפרגמטיקה כוללות שימושים לא־מילוליים בשפה, כגון אירוניה, מטאפורה וכו'. תופעות כאלו הן קשות ביותר לעיבוד ממוחשב.

סקרנו לעיל את היישומים העיקריים של עיבוד שפות טבעיות ואת הקשיים הכרוכים בהם. אין תימה, איפוא, שמצב המחקר הנוכחי אינו מספיק כדי לתת מענה הולם לכל הבעיות המתעוררות. למרות זאת, בעיות רבות הכרוכות בעיבוד ממוחשב של שפות טבעיות נפתרו לחלוטין, ואחרות נפתרו באופן חלקי, עבור שפות בהן הושקע מאמץ רב. לפיכך, שפות כמו אנגלית, גרמנית, או יפנית, אך במידה רבה גם שפות בעלות חוג משתמשים מצומצם יותר, כגון הולנדית, נהנות מיישומים רבים הכוללים טכנולוגיה עדכנית של בלשנות חישובית.

השפה העברית, לעומת זאת, אינה נכללת ברשימה זו. כמובן, העובדה שבעברית משתמש רק מספר קטן יחסית של דוברים הופכת אותה לאטרקטיבית פחות עבור מפתחי טכנולוגיות חדשות. כפי שהראינו לעיל, יישומים רבים בבלשנות חישובית מצריכים ידע מעמיק של השפה, ולפיכך הסבה של טכנולוגיה קיימת משפה אחת לרעותה דורשת משאבים מרובים. הוסף על

כך את העובדה שבעיות רבות הן ייחודיות לעברית (ולשפות שמיות אחרות, שככלל נחקרו פחות), והתוצאה היא שמצב הבלשנות החישובית לעברית, ובייחוד היישומים המעשיים שלה, בכי רע.

נסקור להלן את הקשיים המיוחדים שמציבה העברית בפני מפתחי יישומים בבלשנות חישובית. נוסף על הקשיים האינהרנטיים הנובעים ממורכבות הכושר הלשוני האנושי, העברית מוסיפה שני נדבכים נוספים לקשיים: הכתב והמורפולוגיה.

הכתב העברי

שני קשיים מציב הכתב העברי בפני מפתחי תוכניות לעיבוד ממוחשב של שפה: ראשית, הוא שונה מן הכתב הלטיני, שעבורו פותחו רוב הטכנולוגיות ורוב הכלים החישוביים הקיימים; שנית, הוא נכתב מימין לשמאל, ובכך מונע הסבה פשוטה של יישומים קיימים שפותחו עבור כתבים הנכתבים משמאל לימין.

האלפבית השונה מצריך התייחסות מיוחדת אפילו ביישומים פשוטים יחסית, כגון מנועי חיפוש באינטרנט או תוכנות הגהה אוטומטית. כמובן, לשפות אירופיות רבות אלפבית שונה מזה המשמש את האנגלית, אם בתוספת מספר קטן של סימנים דיאקריטיים (כגון האותיות המציינות את תנועות האומלאוט בגרמנית), ואם לחלוטין (למשל, רוסית או יוונית). לפיכך, כל תוכנה המתוכננת לעבוד באלפבית אחד דורשת הסבה והתאמה לאלפבית אחר. אולם בעיית הכתב העברי היא חמורה יותר, בעיקר משום שהכתב כה חסר. על חסרונות הכתב העברי עומד אורנן (1; 3; 5; 76; 78)[1] והעיקרי שבהם הוא שחלק גדול מן האינפורמציה הטמונה במילים אינו מובע בכתב: היינו, רוב התנועות אינן מיוצגות בכתב העברי חסר הניקוד, שהוא הכתב הנפוץ. מאידך, חלק מהסימנים משמשים לצרכים שונים, למשל כתנועות וכעיצורים. בעיה נוספת היא שמורפמות המבוטאות כמילים עצמאיות בשפות רבות, כגון מילות יחס, מילות חיבור וקישור, מילות שעבוד ותוויות, מתבטאות כצורנים החבורים למילים בעברית.

כפועל יוצא מכל אלה, דרגת ריבוי המשמעות המורפולוגית בעברית גבוהה במיוחד. על־פי אורנן (1) תכונות הכתב העברי מביאות לידי כך שכחמישים אחוזים מתבניות המילים העבריות הן הומוגרפים, כלומר משתייכים ליותר מערך מילוני אחד, ובממוצע מתקבלות ארבע ויותר אפשרויות קריאה לכל

1 המספרים מסמנים פריטים ביבליוגרפיים; ר׳ הרשימה בסוף המאמר.

מילה. לעומתו, טוענות בנטור, אנג'ל ושגב (10) כי קרוב לשישים אחוזים מהמילים מקבלות יותר מניתוח אחד, ושליש מהן יותר משניים, בעוד שלשלוש מאות מילים נפוצות בשפה התקבל ממוצע של 2.7 ניתוחים למילה. בבחינה של טקסטים גדולים, שנבחרו מתוך העיתונות והכילו 40,000 תמניות מילים, נמצא כי מספר הניתוחים הממוצע למילה הוא 2.1, וחמישים וחמישה אחוזים מן המילים קיבלו יותר מניתוח בודד (66; 90).

כאמצעי להתגבר על ריבוי המשמעות, הנגרם עקב בעיות הכתב, הציע אורנן כתב פונמי, המשתמש באותיות האלפבית הלטיני ובמספר סימנים נוספים הנמצאים על מקלדת המחשב (75; 76). הכתב התקבל כתקן, הן ישראלי והן בינלאומי, וכן פותחו תוכניות מחשב הממירות את הכתב הפונמי לכתב עברי ולהיפך (אם כי הכיוון ההפוך נתקל, כמובן, בבעיית ריבוי המשמעות האינהרנטית).

מורפולוגיה עברית

גורם נוסף לקשיים המיוחדים בעיבוד ממוחשב של עברית הוא המורפולוגיה של השפה. תהליכי הנטייה בעברית מבוססים בעיקרם על שרשור צורנים, כמו בשפות אירופיות רבות. אלא שבעברית צורני רישא וצורני סיפא, ולעתים מצטרפים שני סוגי הצורנים לבסיס אחד (כגון בנטיית הפועל ת+שמר+ו). יתר על כן, בעברית משתנה צורת הבסיס בנטיות השונות. כך, למשל, נוח לראות את צורת העבר של הפועל כבסיס הנטייה, אך בזמן עתיד משתנה הבסיס הזה (למשל, שמר הופך לשמור, כגון אשמור, תשמור וכו').

אולם בעיות הנטייה מתגמדות אל מול הבעיות שמציבים תהליכי הגזירה בעברית, ובעיקר מנגנון יצירת המילים המבוסס על שילוב של שורש ומשקל. כאן התהליכים אינם ניתנים לתיאור באמצעות שרשור בלבד, ונדרשים מנגנונים מורכבים יותר על מנת לאפיין אותם. בפרט, בעוד שתהליכי שרשור ניתנים לתיאור (ולמימוש חישובי) באמצעות ביטויים רגולריים ומכונות מצבים סופיות (ראה להלן, פרק 4), ובכך מובטח מימוש יעיל של התהליכים, קשה יותר (אם כי לא בלתי אפשרי) לתאר באמצעים כאלה שילובי שורש-משקל.

3 עבר

בפרק זה נסקרות מערכות ממוחשבות קיימות לעיבוד לשוני בעברית. למרות הקשיים, ולמרות העניין המסחרי המועט יחסית בשפה העברית, הוקדש מאמץ

רב לפיתוח יישומים לעברית, בעיקר באקדמיה אך גם בתעשייה. עקב הקשיים המרובים שמציב עיבוד סמנטי ופרגמטי, מתמקדות מערכות קיימות ברבדים הפונולוגיים והמורפולוגיים של השפה, ולעתים גם בתחביר. מכיוון שלא הוקמה עד היום תשתית חישובית הולמת לפתרון בעיות יסוד — מאגרי לשון, מילונים ממוחשבים, מנתחים מורפולוגיים, תוכנות להפגת עמימות מורפולוגית, לקביעת חלק דיבר ולקביעת משמעות מילונית — שתעמוד לרשות החוקרים באופן חופשי, קשה לבנות מערכות מורכבות יותר, כגון מערכת לתרגום אוטומטי, שכן אלו מניחות בדרך כלל את קיומה של תשתית כזו.

רבות מן העבודות העוסקות בעברית התפרסמו בכתב־העת בלשנות עברית חפשי"ת, שיצא לאור בהוצאת "הועדה לבלשנות חישובית של אוניברסיטת בר אילן" החל מתשרי תש"ל, והמשיך להתפרסם עד שנת 1989, שאז הוחלף בכתב־העת בלשנות עברית המתפרסם עד היום. למרבה הצער, עם השנים קטן חלקם היחסי של המאמרים הדנים בבלשנות חישובית ועלה שיעור המאמרים הדנים בבלשנות פורמלית ושימושית. כבר בגיליון הראשון הופיע מאמר שדן באלגוריתם ליצירת מילים בעברית (81). אלגוריתם מקביל לניתוח מורפולוגי פורסם על ידי Price שנתיים מאוחר יותר (83). בגיליון השני הופיע תקציר של עבודת דוקטורט, כפי הנראה העבודה הראשונה שדנה בתרגום אוטומטי מעברית לאנגלית (82), וכן תואר אלגוריתם לזיהוי השורש של מילים בעברית (65). אפילו שיטה אוטומטית לזיהוי כתב יד בעברית תוארה באחד הגיליונות הראשונים, אם כי לא תואר שם כל מימוש שלה (96). מקור נוסף לעבודות מחקר הוא הקובץ בלשנות חישובית עברית (7), בו קובצו מאמרים שהוצגו בימי עיון שנערכו בשנים 1988־1990 על ידי משרד המדע והטכנולוגיה. כהערת אגב נציין שכנסים אלה התחדשו לאחרונה ומוצגות בהם עבודות הנערכות בישראל, עם דגש על עבודות העוסקות בעברית ובערבית (101). הסקירה שלהלן מסודרת על־פי נושאים, ובכל נושא המיון הוא כרונולוגי. למיטב ידיעתנו, הסקירה כוללת את כל העבודות הדנות בעיבוד ממוחשב של עברית שהתפרסמו עד היום.

מורפולוגיה

אורנן (1) מתאר את מפעל המילון ההיסטורי של האקדמיה ללשון העברית (החל מ־1964), כפרוייקט הממוחשב הרציני הראשון לעיבוד העברית. אולם נראה שפרוייקט זה ניתן לאפיון כשימוש במחשב לצורכי יישום לשוני (הכנת מילון), ללא עיבוד לשוני ממוחשב משמעותי. במסגרת דומה ניתן לכלול גם

עבודות שמטרתן שימוש במחשב לאיסוף אינפורמציה סטטיסטית על פעלים בעברית (71; 72; 73). מן העבר האחר ניצבת עבודתו של אזר (38), המתאר אלגוריתם מלא לניתוחה של כל מילה המופיעה במקרא, שלא נוסה על מחשב.

המערכת הממוחשבת הראשונה לעיבוד עברית נבנתה, כנראה, על ידי שויקה בראשית שנות השישים. המערכת (35) נכתבה בשפת אסמבלר של מחשב "פילקו", שעמד לרשות צה"ל. בשל זמנה, זו מערכת בסיסית ביותר, אולם הישגיה מרשימים: המאמר מתאר הן ניתוח מורפולוגי מלא של מערכת השם (מילון הפעלים לא הושלם בגירסה שפורסמה) והן יישום של הניתוח לצורך הכנת קונקורדנציה ורשימת מובאות. במאמר נוסף (29) מתאר שויקה את החלק המטפל במילים פועליות.

עבודה זו שימשה מבוא לפרוייקט רחב־היקף שעסק בהיבטים שונים של בלשנות חישובית, עיבוד שפות טבעיות ואחזור מידע: פרוייקט השו"ת (שאלות ותשובות), שהחל ב־1967, בראשות אביעזרי פרנקל ואחר כך יעקב שויקה (30; 52; 44). הפרוייקט התבסס על מסד נתונים של 102 כרכים עם שאלות ותשובות הלכתיות שנאספו במהלך כאלף וארבע מאות שנים. מטרת הפרוייקט הייתה אִחזור מידע, בניית קונקורדנציות ומתן שירותי שאילתא שישמשו במסד הנתונים. לצורך כך היה הכרח לפתח כלים לעיבוד לשוני, ובפרט אלגוריתמים לניתוח מורפולוגי (37). פותחו אלגוריתמים ליצירה אוטומטית של כל הצורות הנטויות האפשריות של כל מילת בסיס בעברית, כולל אלו המתקבלות כתוצאה מתוספת מילות יחס, חיבור, שעבוד וכו'. על בסיס אלגוריתם היצירה נבנה קובץ ובו כל המילים העבריות האפשריות, כ־2,500,000 תבניות; בזמן ניתוח מופעלת תוכנית המסירה את המיליות האפשריות ממילת הקלט ובודקת אם הצורה שנותרה נמצאת בקובץ הצורות הנטויות. כך משולבים ניתוח ויצירה מורפולוגיים במערכת שלמה לעיבוד ממוחשב של מילים בעברית (אם כי לא עברית מודרנית).

מאוחר יותר פיתח שויקה בשלישית מערכת לניתוח מורפולוגי של עברית, בשם מלים, במסגרת פרוייקט רב־מלים של המרכז לטכנולוגיה חינוכית (31). מערכת זו הותאמה לעברית מודרנית, ושימשה בסיס לשני יישומים עיקריים: תוכנת ניקוד ומילון שימושי המכיל גם ניתוח מורפולוגי. יישומים אלה הפכו לפיתוחים מסחריים ולפיכך לא הועמדה המערכת, שהייתה כנראה המקיפה ביותר שפותחה לעברית, לרשות החוקרים בבלשנות חישובית. לסקירה קצרה של שלוש העבודות הללו, ראה שויקה (32).

גישה שונה למורפולוגיה של העברית פיתח אורנן. בהסתמך על אבחנותיו הנוגעות למגבלות ולקשיים שמציב הכתב העברי על עיבוד ממוחשב (ראה

לעיל), הציע אורנן את הכתב הפונמי ועל בסיס כתב זה פיתח אורנן מגוון רחב של תוכניות מחשב, ובהן תוכנית לניקוד על ידי מחשב (4), הכנת קונקורדנציות ומפתחות (3), שפותחה במיוחד עבור מסד נתונים של פסקי דין (5), תוכניות לניתוח ויצירה מורפולוגיים (6; 12; 33; 34) וכן תוכניות הממירות כתב עברי רגיל לכתב פונמי (78). לאחרונה שולבו האלגוריתמים הלשוניים שפיתח אורנן במהלך השנים במספר מוצרים מסחריים, ובהם תוכנות עזר להקראת טקסט (לעיוורים וכבדי ראייה, לדיסלקטים ולמרכזנים), מנוע חיפוש רב־לשוני ומערכת אחזור מידע.

כאן ראוי לציין שתי מערכות לניתוח מורפולוגי של עברית, ששולבו במערכות מורכבות יותר, המבצעות ניתוח תחבירי. האחת היא חלק מעבודת הדוקטורט של כהן (18; 19), והניתוח המורפולוגי בה נעשה בשני שלבים: בשלב הראשון מוצאת התוכנית את "הגרעינים" האפשריים של כל מילה, ובשלב השני מיוצרות מתוך הלקסיקון כל הנטיות האפשריות של כל גרעין, ומושוות למילת הקלט. במובן ידוע, זוהי שיטה הדומה לתהליך שתואר לעיל (37). המערכת גם בוחרת את הניתוח ה"סביר" ביותר של כל מילה, על־פי שיקולי הקשר המתוארים ככללי העדפה. שיעור ההצלחה המדווח הוא כחמישים אחוזים, אך אין הערכה מפורטת של התוצאות. המערכת האחרת, HUHU, פותחה באוניברסיטה העברית בירושלים ובה רכיב של ניתוח מורפולוגי שמטרתו להקדים את הניתוח התחבירי (74). שיטת הניתוח מתבססת על מילון ובו, לכל ערך, מידע דקדוקי על כל נטיותיו; וקבוצה של חוקים המתארים את הצורנים בשפה ואופני הצטרפותם לבסיסים השונים. נוסף על כך מטופלות המיליות העשויות להצטרף כרישא למילים באופן נפרד.

ניתוח מורפולוגי הוא פן אחד של מערכת ממוחשבת מסחרית, Context, שמטרתה אחזור מידע מתוך מסמכים (25). מטרת המערכת לשלוף מסמכים (בעברית ובאנגלית) מתוך מסד נתונים מוגדר מראש, לפי דרישה נתונה של המשתמש. לצורך כך הוגדרה "קרבה סמנטית" של מילים המקשרת מילים שונות הקשורות מורפולוגית, פונטית או סמנטית. המערכת אינה משתמשת במילון, ואין במאמר כל דיון בהיקפה או באיכות תוצאותיה.

מערכת מסחרית אחרת, אבגד, פותחה במרכז המדעי של חברת יבמ (10; 11). היא מתבססת על מילון בן 25,000 ערכים, ממנו ניתן לזהות "מאות אלפי" מילים עבריות (על כל נטיותיהן). המילים בלקסיקון נבחרו "מתוך המילים השימושיות בעברית המודרנית". המערכת שולבה במספר מעבדי תמלילים, כששימושה העיקרי הוא תיקון טעויות כתיב. סגל (87) נעזר במערכת זו כדי

לבנות מנתח מורפולוגי חופשי לשימוש: המנתח נבנה תוך יצירת בסיסים אפשריים באופן אוטומטי, הטייתם בכל האופנים האפשריים ובדיקתם מול המנתח הקיים.

הפגת עמימות מורפולוגית (disambiguation)

ברוב המערכות שתוארו בסעיף הקודם מתמצה פעולת המנתח המורפולוגי בהפקת כל הניתוחים האפשריים של המילה. בעברית, כאמור, מספר הניתוחים האפשריים למילה ממוצעת גבוה במיוחד (בהשוואה לשפות אירופיות). לפיכך יש צורך, ביישומים רבים, להפיג את העמימות המורפולוגית: לקבוע, תוך שימוש בשיקולי הקשר קצר וחוקי העדפה, איזה מן הניתוחים האפשריים של מילה מסוימת הוא הסביר ביותר.

ישנם שלושה אופנים עיקריים לשימוש בהקשר לצורך הפגת עמימות. הראשון, שהוא הבטוח ביותר, הוא על ידי איסוף רשימה ממצה של סדרות מילים, שבכל אחת מהן מילה מרובת משמעות, תוך ציון ידני של המשמעות המתאימה של המילה בהקשר בו היא נתונה. פתרון מסוג זה, שבו סדרות המילים הן באורך שתיים, הוצע בעבר (45); ברור כי לשפה כמו עברית הוא כרוך בהשקעת משאבים עצומים, וכנראה אינו ישים. פתרון דומה מוצע גם על ידי כהן (18) ועל ידי לוינגר (22).

אופן אחר הוא חילוץ מידע אודות ההקשר הקצר של מילים מתוך דקדוק נתון. הרעיון העומד ביסוד השיטה הוא שאילו היה באפשרותנו לנתח משפט נתון ניתוח תחבירי, הרי שהניתוח היה מפיג חלק גדול מן העמימות המורפולוגית. אמנם ניתוח תחבירי הוא יקר מבחינה חישובית, אך אם נתון דקדוק המתאר את תחביר השפה, ניתן לייצר ממנו באופן אוטומטי מידע, שעל בסיסו אפשר יהיה להפיג את העמימות במהירות וביעילות. שיטה כזו, עם יישום לתיקון שגיאות בעברית, הוצעה על ידי הרץ ורימון (55 ; 14).

הגישה השלישית מציעה שימוש בשיטות סטטיסטיות. אם נתון מאגר לשון (קורפוס) ובו לכל מילה מצוינים כל ניתוחיה האפשריים ובצדם גם הניתוח הנכון (קורפוס כזה חייב להיות מעשה ידי אדם), ניתן להשתמש בשיטות סטטיסטיות על מנת ללמוד נתונים מתוך המאגר ולהשליכם על דוגמות חדשות, שעדיין לא סומנו. רעיון זה מומש על ידי לוינגר, אורנן ואיתי (66): עיקר העבודה הוא אלגוריתם המחשב באופן אוטומטי קירוב טוב להסתברות המורפו־לקסיקלית של כל מילה בטקסט נתון. ההסתברות המורפו־לקסיקלית של ניתוח מסוים של מילה היא ההסתברות שהניתוח הוא אכן הניתוח הנכון של המילה

(במונחים סטטיסטיים, זוהי ההסתברות המותנית של הניתוח בהינתן המילה). בעבודה מתוארת דרך שיטתית לחישוב הסתברויות אלו מתוך מאגר לא־מתויג, כלומר קורפוס גדול של משפטים לא מעובדים. בהתבסס על ההסתברויות המורפו־לקסיקליות, מציע המאמר לבחור לכל מילה רב משמעית בטקסט את הניתוח שהסתברותו גבוהה ביותר. העבודה מומשה בעזרת המנתח המורפולוגי של יבמ (10), תוצאותיה הוערכו בפירוט ושיעור ההצלחה המדווח גבוה.

פתרון דומה מציע גם סגל (24): הוא מנתח טקסטים תוך שימוש במנתח המורפולוגי שלו שהוזכר לעיל; לאחר מכן נבחר לכל מילה בטקסט ניתוחה הסביר ביותר, על־פי הסתברויות שחושבו מתוך קורפוס מנותח לא גדול. בשלב הבא מתקנת המערכת את החלטותיה תוך הסתמכות על הקשר קצר (מילה אחת מכל צד). גם פקודות התיקון הללו נלמדות באופן אוטומטי מתוך הקורפוס המנותח. בשלב האחרון מתוקן הניתוח על־פי ניתוח תחבירי של כל המשפט. התוצאות המדווחות הן מצוינות, אלא שזמני הביצוע של התוכנית אינם קבילים (זמן הריצה המדווח על "שני מאמרים" הוא שלושים דקות). ללא הניתוח התחבירי, התוצאות הן טובות וזמן הביצוע סביר.

גישה שונה, אף היא מבוססת על שיטות סטטיסטיות, הציעו דגן ואיתי (46). השיטה מתבססת על הבדלים בין מיפוי של מילים למשמעויות בשפות שונות, ומניחה שבשפה אחרת — למשל, אנגלית — קיימים כבר נתונים סטטיסטיים המאפשרים לבחור את הקריאה הנכונה של מילים רב־משמעיות. אם נתון מנתח תחבירי המזהה קשרי תפקידים (כגון נושא־פועל, או פועל־מושא וכדומה) בשפת המקור (כאן, בעברית), ניתן למפות את הקריאות השונות של המילים בכל מופע של קשר כזה לשפה האחרת ולהשתמש בנתונים של השפה האחרת כדי לבחור את הקריאה הנכונה בשפת המקור. השיטה נוסתה על עברית ועל גרמנית, תוך שימוש באנגלית כשפה השנייה, והניבה תוצאות טובות. הבעיה העיקרית בשיטה זו היא התבססותה על מנתח תחבירי, ולו חלקי, לעברית.

תחביר

עד היום טרם נכתב מנתח תחבירי מקיף, המבוסס על מודל תיאורטי מקובל, עבור השפה העברית. הבעיה העיקרית בדרך אל מנתח כזה היא שטרם נכתב דקדוק חישובי מקיף וממצה עבור העברית. ספרי הדקדוק המקובלים, ואף אלו המכוונים לתיאור פורמלי של השפה (2; 16; 26; 27; 28; 53), אינם מספקים לצורך עיבוד ממוחשב.

ניסיון מעניין לנסח דקדוק חישובי עבור השפה העברית נעשה על ידי Price בסוף שנות השישים (84). בעבודה זו מתאר המחבר דקדוק מבנה צירופים של תחביר העברית המודרנית: הוא מנסח 179 חוקים, מהם נבדקו רק 111. מתוארים 26 משפטים שהדקדוק הצליח לנתח, כולם פשוטים וקצרים (לכל היותר 10 מילים) ורובם דומים מאוד זה לזה. העבודה הייתה אמנם חלוצית, אך הישגיה היו זעומים.

ניתוח תחבירי "מכני" (לא ממוחשב) של טקסט בעברית בוצע על ידי אזר (8). העבודה מתארת אלגוריתם לניתוח תחבירי ומפרטת כארבעים חוקים. הדקדוק לוקה ביצירת־יתר: החוקים אינם מעודנים מספיק ויאפשרו יצירת מבנים רבים שאינם חוקיים בשפה. אולם המגרעה העיקרית בעבודה זו היא שהאלגוריתם לא מומש מעולם ועל כן גם לא נבדק באופן מעשי.

המנתח התחבירי הראשון לעברית הוא כפי הנראה HUHU, שפותח על ידי נירנבורג ובן־אשר באוניברסיטה העברית בירושלים (74). המערכת מבוססת על רשתות מעבר מורחבות (ATN) והיא מקבלת כקלט משפטים בעברית לא מנוקדת, מנתחת את המילים ניתוח מורפולוגי ואז מנתחת את מבנה המשפט. קשה לדעת בדיוק מה היה היקף הדקדוק שנכתב למערכת, אולם על־פי פירוט חלקי הדיבר שציינו המחברים, ולפי תיאור של מספר חוקים, נראה שההיקף לא היה רחב. המנתח פעל באופן אי־דטרמיניסטי (כלומר, לא יכול להכריע בין אפשרויות שונות להתקדמות במהלך הניתוח) והפיק את כל הניתוחים האפשריים לכל משפט.

עבודה מוקדמת אחרת היא עבודת הדוקטורט של כהן (18). גם כהן תקף את הבעיה של ניתוח תחבירי במשולב עם ניתוח מורפולוגי, וגם הוא השתמש בכתב עברי לא מנוקד כקלט. הוא מדווח על אחוזי הצלחה גבוהים, אלא שהמנתח שלו נוסה על חמישים משפטים בלבד. מספר החוקים התחביריים בדקדוק שכתב הוא שישים. אחת הבעיות העיקריות במערכת זו היא שהדקדוק משולב כחלק אינטגרלי בתוכנית הניתוח, ולפיכך כל שינוי או עדכון של אחד מחוקי הדקדוק מצריך עדכון של התוכנית כולה.

אלבק (9) מציעה שיטת ניתוח תחבירי חדשנית לעברית. כדי להתגבר על ריבוי המשמעות המורפולוגי ועל סדר המילים החופשי יחסית בעברית, ומתוך התבוננות בתהליך האנושי של הבנת שפה כתובה, ובפרט הפגת העמימות, היא מציעה אלגוריתם ניתוח, שבו המילה מקבלת משמעות יחידה מיד עם קריאתה ללא הצצה קדימה וללא תיקון טעויות לאחור, תוך היעזרות במיגוון חוקים, מהם הכרחיים ומהם חוקי העדפה, המייצרים, מעדכנים ומבטלים ציפיות בנוגע

לתפקיד המילה המסוימת במשפט. המערכת המתוארת כוללת 192 חוקים תחביריים, המספיקים כדי להפיק את הקריאה הנכונה לכ־98 אחוזים מהמילים בטקסט בן מאה משפטים. המאמר מגביל מראש את סוג משפטי הקלט לכאלה שבהם סדר המילים "מקובל", מבנה המשפט "מסודר וקבוע" והמשפטים "מלאים וברורים". כמו כן, אין המאמר מפרט או מדגים את הניתוח התחבירי המוענק למשפטים, אלא רק את הפגת העמימות המורפולוגית. גם נתונים על הביצועים החישוביים של המערכת אינם נידונים.

כמובן, כל ניתוח תחבירי תלוי באופן הדוק בתיאוריה לשונית כלשהי. בשני העשורים האחרונים מסתמנות מספר תיאוריות לשוניות המנוסחות באופן פורמלי יחסית ומתאימות יותר מאחרות למימושים חישוביים. את התיאוריות הללו מאפיין כוח תיאור חישובי (אקספרסיביות) גבוה יחסית, ובכל אופן גבוה יותר מזה של דקדוקים חסרי הקשר. רובן עושות שימוש כזה או אחר בפעולת ההאחדה (unification), ורובן מייצגות את המידע הלשוני — הלקסיקון, הדקדוק, הייצוגים של מבעים ואף תוצאת הניתוח כולה — במבנים מורכבים הקרויים מבני תכוניות (feature structures) (88). בין התיאוריות הללו: LFG — Lexical Functional Grammar (51 ; 58); HPSG — Head-Driven Phrase Structure Grammar (80–79); TAG — Tree-adjoining Grammar (57); CG — Categorial Grammar (54 ; 93). תיאוריות אלו ודומותיהן משמשות מצע למחקרים תיאורטיים בבלשנות, שתוצאותיהם דקדוקים חישוביים המתארים תופעות שונות במיגוון רחב של שפות טבעיות. בהתחשב בפורמליות של תיאוריות אלו, ניתן להתייחס אליהן כאל שפות מחשב עיליות, ואל הדקדוקים המתוארים בתיאוריות אלו כאל תוכניות מחשב. דקדוקים אלו ניתנים לבדיקה ולהרצה באמצעות מחשב. כמכנה משותף כללי ביותר לרבות מהתיאוריות הלשוניות הללו ניתן לבחור בפורמליזם הנקרא PATR, שאמנם אינו תיאוריה לשונית, אך מאפשר ניסוח של דקדוקים המבוססים על מבני תכוניות והאחדה באופן שעולה בקנה אחד עם התיאוריות השונות (89 ; 88).

דקדוק חישובי לשפה העברית העומד בפני עצמו (כלומר, אינו משולב בתוכנת ניתוח) נכתב לראשונה על ידי וינטנר (98). העבודה כללה סקירה של מספר פורמליזמים לשוניים והתאמתם לעברית (102), דקדוק קטן ב־PATR (15) ודקדוק מקיף יותר, המבוסס על עקרונות התיאוריה הלשונית LFG, אם כי אינו צמוד אליה בכל פרטיה (98 ; 103 ; 104). הדקדוק מפיק תיאור כפול: הן של המבנה הפורמלי של משפטי הקלט (במונחים של דקדוקים גנרטיביים) והן של המבנה הפונקציונלי שלהם. הוא נבדק על מאגר קטן של משפטים

בעברית קלה, שנלקחו מתוך העיתון שער, ושיעור ההצלחה על קורפוס זה הוא שבעים וחמישה אחוזים.

במסגרת דומה עבדה גם יצהר, שהתמקדה בצירופים שמניים בעברית (17). העבודה מספקת תיאור מפורט ומקיף של צירופים כאלה, בגישה המושפעת מ־LFG, והדקדוק מומש ונבדק. עבודות אחרות העוסקות בתחביר העברית החדשה בגישה חישובית, לפי שיטת התיאוריה הלשונית HPSG, כוללות לקסיקון חישובי של מערכת הפועל העברי (91); דקדוק חישובי קטן לעברית (99) והרחבתו לתיאור של צירופים שמניים (100); ותיאור של מבנה צירופי זיקה (95).

דקדוקים חישוביים הנכתבים בסביבות מבוססות האחדה ניתנים לשימוש, באופן תיאורטי, הן לניתוח והן ליצירה, אך הבעיות הכרוכות ביצירת משפטים הן קשות וסבוכות, ובאופן מעשי דקדוקים המתוכננים לצורך ניתוח לא יוכלו לשמש ליצירה. את הבעיות הכרוכות ביצירת משפטים בעברית תוקפים דהן־נצר ואלחדד בסדרה של עבודות. דהן־נצר הסבה את המערכת שפותחה על ידי אלחדד ליצירת צירופים שמניים בעברית (47). בעבודות אחרות הם עוסקים בבעיות הספציפיות הכרוכות ביצירת צירופי סמיכות (49), כמתים (48) וצירופי יחס (50).

עבודות אחרות

תוכנית ראשונה ליצירת דיבור על ידי מחשב מתוארת אצל לאופר (20). מדובר במערכת היוצרת דיבור מלאכותי מקלט הנתון בכתב פונטי (כולל מקום הטעם), תוך הפעלת חוקי מעבר (המממשים קשרי גומלין בין הגאים סמוכים) וחוקי הנגנה. מעניין לציין כי בימים אלה ממש מפותחת במרכז המחקר של חברת יבמ בחיפה מערכת לזיהוי דיבור בעברית. כמובן, בעיית הזיהוי קשה בהרבה מבעיית היצירה.

מעטות העבודות העוסקות באופן ישיר בסמנטיקה של עברית. סמואלסדורף (86) מתאר ניתוח סמנטי לצורך תרגום ממוחשב, אך העבודה אינה מבוססת על כל תיאוריה שהיא ואין בה תיאור ביצועים מעבר לרמת המילה הבודדת. ברמת המילה, מציע ניסן (23) מערכת ליצירת מונחים עבריים, אך חשיבותה המעשית לא ברורה. בשקנסקי ואורנן מציעים מערכת כלים לסיוע בתרגום אוטומטי, עם יישום לתרגום מעברית לרוסית (39). מערכת לתרגום אוטומטי תוך שימוש בתכונות סמנטיות, המודגמת בתרגום מעברית לאנגלית ולספרדית, מתוארת על ידי אורנן וגוטר (77).

4 עתיד

כפי שהראינו בפרק הקודם, מיגוון המערכות הממוחשבות לעיבוד עברית הוא רחב, אך עדיין חסרה תשתית רחבה, מבוססת ונגישה, שתאפשר פיתוח תוכנה מודרנית וביסוס מחקר מתקדם בבלשנות חישובית עברית. בפרק זה נסקרות הדרישות מתשתית חישובית כזו.

כפי שצוין לעיל, כמעט כל יישום המצריך ידע לשוני נזקק לאוצר מילים מלא של השפה. שפות טבעיות הן דינמיות: לא ניתן לקבוע ברגע נתון מהו אוצר המילים של שפה כלשהי. הסיבה העיקרית לכך היא ההצטרפות המתמדת של מילים זרות, ובעיקר שמות פרטיים, לשפה, ושכיחותן הגבוהה של מילים אלו בטקסטים נפוצים (כגון מאמרי עיתון או מסמכי אינטרנט). יתרה מזו, מילים זרות (ואף שמות) משתלבות לעתים קרובות במרקם של השפה, ואף מצייתות במידה מסוימת לחוקי המורפולוגיה שלה. כך, למשל, נוסף לשפה העברית שם העצם כלנתריזם, הנגזר מהשם הפרטי כלנתר; ושם העצם טלפון משמש בסיס לפועל טלפן על כל נטיותיו. לקסיקונים ממוחשבים חייבים להיות רגישים לדינמיות של השפה, ולפיכך מן הראוי להשקיע מאמץ בפיתוח טכנולוגיה שתעשיר לקסיקונים קיימים במילים חדשות, תוך שימוש בידע הנרכש על ידי סריקה מתמדת של מסמכים (למשל, ניתן לפתח תוכנה שתסרוק באופן קבוע מסמכי אינטרנט בעברית ותפיק רשימה של מילים לא מוכרות).

אולם הלקסיקון הוא רק השלב הראשון בעיבוד ממוחשב של מילים; בשפה כמו עברית, בעלת מורפולוגיה עשירה יחסית, תהיה זו טעות לשמור לקסיקון של כל הצורות הנטויות של כל הבסיסים. חוקי המורפולוגיה של העברית הם פשוטים יחסית, ומן הראוי לייצגם באופן מדויק ופורמלי כדי לאפשר בנייה של מנתחים מורפולוגיים מודרניים, שניתן להרחיב, לשפר ולתחזק אותם עם התרחבות הלקסיקון כמתואר לעיל. רוב העבודות שתוארו בפרק הקודם מתייחסות לבעיית המורפולוגיה העברית כאל בעיה שונה במהותה מהמורפולוגיה של שפות אחרות. במשך שנים ארוכות הוזנח תחום המורפולוגיה החישובית בעולם, עקב העיסוק המוגבר בשפה האנגלית, שלה מורפולוגיה מנוונת. אלא שמשנות השמונים החל להתפתח בעולם תחום מחקר ייחודי שעוסק בגישות חישוביות למורפולוגיה (ופונולוגיה) של שפות טבעיות, גישות המתבססות על מכונות מצבים סופיות (finite-state machines) (92 ; 59 ; 85). חלוץ הגישה הזאת היה קוסקניימי, שפיתח מודל חישובי לתיאור תהליכים מורפולוגיים ופונולוגיים באמצעים כאלה, שנודע בשם מודל שתי־הרמות (62).

הניסיון הראשון לבדוק את התאמת המודל הזה לעברית נערך על ידי לביא, ובמהלכו מומש מחדש המודל בשפת פרולוג (63; 64; 21). מסקנות הניסוי הן כי בעוד שהמודל מתאים לתיאור תהליכי הנטייה של הפועל, הרי שתהליכי גזירה מסוימים, כגון יצירת בסיסי משנה מהבסיס הראשוני, קשים מאוד להבעה במודל זה.

אולם בשנים האחרונות חלו התפתחויות חשובות בתחום זה, והוצעו הרחבות של מודלים המבוססים על טכנולוגיה של מכונות מצבים סופיות המאפשרות עושר ביטוי רב יותר (60; 69; 42; 97; 70). הרחבות אלו מאפשרות לנסח את התהליכים המורפולוגיים בשפת תיאור נוחה, ולהפיק באופן אוטומטי מנתחים מורפולוגיים, ובדרך כלל גם תוכניות יצירה מורפולוגיות. שפות תיאור כאלה מבוססות על ביטויים רגולריים (regular expressions) והן מאפשרות לבלשן לאפיין את המבנה המורפו־פונולוגי של מילים בשפה תוך שימוש בשני סוגי סימנים: האלפבית של השפה וקבוצה של סימנים מיוחדים, המציינים פעולות על סדרות של סימני אלפבית (מחרוזות). פעולות כאלו כוללות, למשל, שרשור: צירוף סדרתי של שתי מחרוזות למחרוזת אחת; אך במודלים המורחבים ניתן להשתמש בפעולות מורכבות יותר, ובהן גם קבוצה גדולה של פעולות החלפה על מחרוזות, בדומה לחוקי ההחלפה המוכרים מהגישה הגנרטיבית לפונולוגיה. למשל, ניתן לבטא בשפות אלו בקלות מרובה את החוק המחייב חילוף של ה־ת של בניין התפעל בעיצור הראשון של השורש, כאשר עיצור זה הוא שורק. מאפיין חשוב של מודלים אלה הוא שניתן לממשם במחשב ביעילות מרבית. ואמנם, המערכות שצוינו לעיל כוללות תוכנה המתרגמת את הביטויים הרגולריים המורחבים לתוכניות מחשב יעילות ביותר.

ואכן, המודלים המורחבים שימשו ליצירת מנתחים מורפולוגיים מודרניים למספר שפות שמיות, ובהן ערבית (40; 41; 61). מן הראוי להשתמש בטכנולוגיה מודרנית זו על מנת לפתח מנתחים מורפולוגיים לעברית (על כל משלביה) שיהיו פתוחים לשימוש לכל צורך ויתוחזקו באופן שוטף כך שיישארו מעודכנים תמיד.

השיטות להפגת עמימות מורפולוגית, תוך שימוש בהקשר קצר, שתוארו בפרק הקודם הן מבטיחות, ומן הראוי להמשיך ולפתחן כחלק ממנתח מורפולוגי מודרני. על בסיס תוצאות אלו ניתן יהיה לפתח מנתחים תחביריים שטחיים (shallow parsers), אשר מפיקים שלד של המבנה התחבירי של כל משפט בזמן קצר בהרבה מהזמן הנדרש לניתוח תחבירי מלא. גם לצורך יישומים כאלה ניתן להשתמש בטכנולוגיה של מכונות מצבים סופיות. עבודות כאלו נערכות עתה

באוניברסיטת בן גוריון, ומטרתן להסב שיטות קיימות של תיוג חלקי־דיבר (part-of-speech tagging) לעברית (36). תיוג, משמעותו סימון חלק הדיבר הנכון לכל מילה המופיעה בקורפוס, ולעתים ניתן לסמן גם מידע נוסף על חלק הדיבר, כגון תכונות מורפולוגיות ותחביריות או משמעות מילונית (word sense).

יתרון גדול של עבודות כאלו הוא בהתאמתן לשפות שמיות אחרות. המורפולוגיה של העברית קרובה מאוד לזו של שפות שמיות אחרות, וביניהן חשובה במיוחד הערבית. לפיכך, כלים ושיטות שיפותחו עבור העברית, סביר מאוד שיוכלו לשמש גם לצורך מחקר לשוני ופיתוח מערכות לעיבוד הערבית. אמנם, קיימות מערכות רבות העוסקות בשפה הערבית, ויקצר כאן המצע מלפרט את כולן; ובכל זאת, נראה כי יש טעם להתמקד בפיתוח שיטות שיאפשרו יישומים בשתי השפות. למשל, בפרוייקט שהחל לאחרונה באוניברסיטת חיפה, מנסים לפתח טכנולוגיה שתאפשר לימוד אוטומטי של לקסיקון ומנתח מורפולוגי של ערבית ישראלית מדוברת מתוך קורפוס נתון (94). יש יסוד לקוות שטכנולוגיה זאת תאפשר גם לימוד של לקסיקון וחוקים מורפולוגיים לעברית.

בנוסף ליישומים לשוניים מובהקים כמו לקסיקון ומנתח מורפולוגי, הניסיון מלמד כי יש צורך להשקיע בפיתוח כלים שונים שיסייעו לבניית מערכות ממוחשבות לעיבוד שפה. לגבי מיגוון רחב של יישומים, מסתבר כי גישות שאינן משלבות ידע לשוני מעמיק, אלא משתמשות בידע סטטיסטי על השפה, מגיעות לביצועים מרחיקי לכת. ביישומים מסוימים, כגון זיהוי דיבור (speech recognition), מצליחות גישות כאלו להכות שיטות המבוססות על ידע לשוני שוק על ירך. אלא שלצורך איסוף המידע הסטטיסטי, ובעיקר לשם אימון המערכת, יש צורך במאגרי מידע גדולים. מאגרים כאלה, קורפוסים של מישלבים שונים של שפה (הן כתובה והן מדוברת), הם הכרחיים לצורך עיבוד סטטיסטי. יתרה מזו, לצורך אימון מערכות כאלו נהוג פעמים רבות לתייג חלק ניכר מהמאגר. בשפה כמו עברית, ריבוי המשמעות המורפולוגי הגבוה יצריך, כפי הנראה, גם סימון הניתוח המורפולוגי הנכון של כל מילה בחלק הקורפוס המתויג. יתרה מזו, לשפות שנחקרו היטב, כגון אנגלית או גרמנית, קיימים קורפוסים שחלקים שלמים מהם נותחו באופן תחבירי, ולכל משפט הוצמד עץ גזירה המתאר את מבנהו. קורפוס כזה מכונה **בנק עצים** (tree bank) וחשיבותו לפיתוח יישומים המבוססים על ידע לשוני שטחי לא תסולא בפז. אלא שהכנה של קורפוסים מתויגים היא תהליך ארוך, קשה ומייגע, המצריך בעיקר עבודת

נמלים של בלשנים חישובייים, שלהם ידע מספיק בשפה וידע שטחי לפחות בשימושי מחשב. אין ספק שיש צורך להשקיע משאבים בפיתוח קורפוסים כאלה לעברית.

שתי עבודות נוכחיות עוסקות בכינון מאגרים לשוניים של השפה העברית. מכיוון ששתיהן טרם הושלמו, לא ניתן להעריך כאן את תוצאותיהן, אך חשוב לתאר את מטרותיהן המוצהרות. העבודה הראשונה עוסקת בהקמת בנק עצים לעברית (90). בשלב הראשון, נבחרה קבוצה של 500 משפטים, שנותחו מורפולוגית באופן אוטומטי ולאחר מכן תוקנו תוצאות הניתוח באופן ידני. נוסף על כך נותחו כל המשפטים ניתוח תחבירי (ידנית), ולכל אחד מהם הותאם עץ גזירה המתאר את מבנהו. מטרת העבודה היא לפתח פרוצדורה אוטומטית למחצה שתאפשר הרחבה של בנק העצים בעתיד.

קורפוס אחר עוסק בעברית מדוברת (56; 13): מטרת העבודה היא לכונן מאגר הקלטות מייצג של העברית המדוברת בישראל כיום, שישמש לצורכי מחקר בתחומים מגוונים הקשורים לשפה העברית. הקורפוס ילווה בתעתיקים של ההקלטות, ובכלים שיאפשרו חיפושים, השוואות וניתוחים הן של הדיבור המוקלט והן של התעתיקים. כאמור, פרוייקט זה נמצא עדיין בשלבי תכנון ראשוניים.

ליישומים רבים יספיקו לקסיקון טוב, רחב־היקף, המלווה במנתח מורפולוגי יעיל ומשתמש בשיקולי הקשר קצר לצורך הפגת עמימות מורפולוגית. למשל, השיטות המוצלחות ביותר כיום לתמצות מסמכים (68; 67) מתבססות על ידע לשוני שטחי בלבד, ואינן מצריכות ניתוח מעמיק יותר. כך גם ביישומים של חיפוש אינטליגנטי במאגרי מידע, מיצוי מידע וכדומה. אולם, לצורך יישומים מעמיקים יותר, כגון מענה על שאלות וכיו"ב, יש צורך בהבנת המבנה של מבעים במילים אחרות, וכן יש צורך בניתוח תחבירי.

כאמור, מעטים הדקדוקים החישובייים שפותחו לעברית. על מנת לפתח יישומי שפה טבעית שיכללו הבנה מעמיקה של משמעות המבעים בעברית, יש צורך להתמקד בפיתוח מואץ ורחב־היקף של דקדוקים כאלה. הוכח בעבר שלצורך יישומים מורכבים, כגון תרגום אוטומטי, ביצועיהן של מערכות המשלבות ידע לשוני מעמיק עולים לאין שיעור על מערכות המתבססות על מידע סטטיסטי בלבד. עבודה כזו מצריכה מאמץ רב ושיתוף פעולה בין מספר גדול של חוקרים, בלשנים תיאורטיים, בלשנים חישובייים ומדעני מחשב, אך שכרה בצדה.

כאשר יהיה ברשותנו דקדוק חישובי רחב־היקף של עברית מודרנית, מבוסס על מנתח מורפולוגי מפיג עמימות ומסוגל להתמודד עם קלטים מורכבים, הן

חוקיים והן לא־חוקיים, נוכל להתמודד עם האתגר הגדול ביותר של עיבוד שפות טבעיות: הבנת המשמעות של מבעים, החל מצירופים פשוטים, עבור דרך משפטים וכלה במבני שיח (discourse) מורכבים. אולם אתגר זה יהיה קל יותר מהקודמים לו. סביר להניח שההתמודדות עם הסמנטיקה של העברית לא תצריך כלים, טכנולוגיה ופיתוחים שונים מאלה שכבר קיימים עבור שפות אחרות. זאת משום שהעברית שונה מלשונות המערב בעיקר במורפולוגיה ובתחביר, אולם נדמה שהסמנטיקה הלקסיקלית מציבה פחות שונות בין הלשונות הללו.

ביום שבו נוכל להציג מערכת לעיבוד שפה בעברית שתכיל את כל המרכיבים המתוארים לעיל, תיפתח בפנינו הדרך למיגוון רחב של יישומים כמו גם לפתרון בעיות מחקר מרתקות שכיום נבצר מאיתנו להתמודד איתן. אז תהיה העברית לחברה שוות זכויות במשפחת השפות הניתנות לעיבוד חישובי. אין צורך לומר שפיתוחים כאלה לא יוכלו לעולם להיות מונעים על ידי צורכי תעשיית התוכנה, שכן שוק התוכנות לעיבוד עברית יישאר לעולם מוגבל ביותר. פיתוחים כאלה, הן במסגרת של מחקר בסיסי והן במסגרת של מחקר יישומי, חייבים להיות מונעים על ידי מי שעתיד השפה העברית, בעידן של גלובליזציה מואצת, קרוב ללבו.

תודות

תודתי נתונה למורי רימון, לעוזי אורנן ולאלון איתי שקראו גרסאות קודמות של מאמר זה והעירו עליהן. תודה מיוחדת לשלמה יזרעאל על העידוד ועל הערות רבות ומועילות. טעויות, אי־דיוקים ותפיסות שגויות הם, כמובן, על אחריותי הבלעדית.

ביבליוגרפיה

(1) אורנן, עוזי. תשל״ז. דיווח על מחקר לשוני במחשב המבוצע בישראל. **בלשנות עברית** חפ״שית 11: 121–127.

(2) אורנן, עוזי. תשל״ט. **המשפט הפשוט**. ירושלים: אקדמון.

(3) אורנן, עוזי. תשמ״ה. מפתחות וקונקורדנציות בכתב עברי פונמי. **דברי ימי הקונגרס ה־9 למדעי היהדות**, ירושלים.

(4) אורנן, עוזי. תשמ״ה. ניקוד ע״י מחשב: לקח בלשני. בתוך: ב״צ לוריא (עורך). **ספר אברהם אבן־שושן**. ירושלים: קרית־ספר.

(5) אורנן, עוזי. תשמ״ח. עיבוד טקסטים עבריים במחשב על יסוד כתב חד־משמעי. **משפטים** י״ז: 15–24.

(6) אורנן, עוזי וודים קזצקי. 1986. תהליכי ניתוח ויצירה במורפולוגיה העברית. הכנס הארצי ה־21 לעיבוד נתונים, אילי"א. 153–164.

(7) אורנן, עוזי, גדעון אריאלי ועידית דורון (עורכים). 1992. בלשנות חישובית עברית: קובץ מאמרים מימי עיון שנערכו בשנים 1988, 1989, 1990 ע"י משרד המדע והטכנולוגיה.

(8) אזר, משה. 1972. ניתוח תחבירי מכני: השיטה והפעלתה על ספר רות. בלשנות עברית חפ"שית 5: 1–50.

(9) אלבק, אורלי. 1995. מפצ"ח מבנים: שיטה לניתוח פורמאלי של משפט עברי לא מנוקד. בלשנות עברית 39.

(10) בנטור, אסתר, אביאלה אנג'ל ודנית שגב. תשנ"ג. ניתוח ממוחשב של מלים עבריות. בלשנות עברית 36: 33–37.

(11) בנטור, אסתר, אביאלה אנג'ל, דנית בן ארי־שגב ואלון לביא. 1992. ניתוח ממוחשב של מילים עבריות. בתוך: עוזי אורנן, גדעון אריאלי ועידית דורון (עורכים). בלשנות חישובית עברית: 36–38.

(12) גולדשטיין, ליאור. 1991. ניתוח ויצירה של נטיית הקנין של שמות מלרעיים. חיבור לקבלת תואר מגיסטר למדעים, הטכניון.

(13) הרי, בנימין ושלמה יזרעאל. המודל התכנוני של מאגר העברית המדוברת בישראל (מעמ"ד). (בכרך זה).

(14) הרץ, יצחק ומורי רימון. 1992. מיתון עמימות לקסיקלית ושימושים נוספים של אוטומט הקשר קצר. בתוך: עוזי אורנן, גדעון אריאלי ועידית דורון (עורכים). בלשנות חישובית עברית: 74–87.

(15) וינטנר, שולי. 1992. ניתוח תחבירי של משפטים בעברית באמצעות PATR. בתוך: עוזי אורנן, גדעון אריאלי ועידית דורון (עורכים). בלשנות חישובית עברית: 105–115.

(16) חן, משה וזאב דרור. 1976. מבוא לדקדוק תצרוני עברי. מפעלים אוניברסיטאיים להוצאה לאור.

(17) יצהר, דנה. 1993. דקדוק חישובי לצרופים שמניים בעברית. עבודת מוסמך, מחלקת מדעי המחשב, האוניברסיטה העברית.

(18) כהן, דניאל. תשמ"ד. ניתוח תחבירי מכני של משפט בעברית. חיבור לקבלת תואר דוקטור לפילוסופיה, האוניברסיטה העברית, ירושלים.

(19) כהן, דניאל. תשמ"ה. ניתוח טקסטים לא־מנוקדים וניקודם על־ידי מחשב. דברי ימי הקונגרס ה־9 למדעי היהדות, ירושלים. 117–122.

(20) לאופר, אשר. תשל"ו. דיבור עברי מלאכותי בעזרת מחשב. לשוננו מ: 67–78.

(21) לביא, אלון. 1989. שיטת שתי הרמות למורפולוגיה עברית. חיבור לקבלת תואר מגיסטר, הטכניון.

(22) לוינגר, משה. 1992. הפגת עמימות מורפולוגית בעברית. חיבור לקבלת תואר מגיסטר למדעים, הטכניון.

(23) ניסן, אפרים. תשנ"ג. המטבעה הלשונית. בלשנות עברית 36: 39–49.

(24) סגל, אראל. תש"ס. מנתח צורני הסתברותי לטקסטים עבריים לא מנוקדים. חיבור לקבלת תואר מגיסטר למדעים, הטכניון.

(25) פנקס, גדי. תשמ"ו. מערכת לינגואיסטית לאחזור מידע. מעשה חושב 12: 10–16.

(26) רוזן, חיים. 1967. עברית טובה, עיונים בתחביר. ירושלים: קרית־ספר.

(27) רובינשטיין, אליעזר. תשכ"ט. המשפט השמני, עיונים בתחביר ימינו. הוצאת הקיבוץ המאוחד והאוניברסיטה של תל אביב.

(28) רובינשטיין, אליעזר. תשל"א. הצירוף הפועלי. הוצאת הקיבוץ המאוחד.

(29) שויקה, יעקב. 1965. מחשבים ודקדוק: ניתוח מכני של הפועל העברי. קובץ ההרצאות של הכנס הארצי לעיבוד נתונים. האיגוד הישראלי לעיבוד אינפורמציה: 49–66.

(30) שויקה, יעקב. 1972. טכניקות מהירות לאיתור ושליפה במלון ובקונקורדנציה בעברית. בלשנות עברית חפ"שית 6: 12–32.

(31) שויקה, יעקב. 1990. מלי"ם — מערכת לניתוח דקדוקי מלא, מדויק ומקוון למורפולוגיה של העברית בת זמננו בסביבת מחשב אישי ובסביבת ווקס. הכנס השנתי על מחשבים בחינוך. 63.

(32) שויקה, יעקב. תשנ"ד. הערות למאמר "ניתוח ממוחשב של מלים עבריות" שנדפס בבלשנות עברית 36. בלשנות עברית 37: 87.

(33) שני־קליין, מיכל. 1990. ניתוח ויצירה של נטיית השמות הסגוליים בשפה העברית. חיבור לקבלת תואר מגיסטר למדעים, הטכניון.

(34) שני־קליין, מיכל ועוזי אורנן. 1992. ניתוח ויצירה של נטיית השמות הסגוליים בשפה העברית. בתוך: עוזי אורנן, גדעון אריאלי ועידית דורון (עורכים). בלשנות חישובית עברית.

(35) שפירא, מאיר ויעקב שויקה. תשכ"ד. ניתוח מיכאנוגראפי של המורפולוגיה העברית: אפשרויות והישגים. לשוננו 4/28: 354–372.

(36) Adler, Meni and Miki Tebeka. 2001. Unsupervised Hebrew Part-of-Speech Tagging. In: Shuly Wintner (ed.). *Israeli Seminar on Computational Linguistics (ISCOL'01)*. 19–20.

(37) Attar, R., Y. Choueka, N. Dershowitz and A. S. Fraenkel. 1978. KEDMA — Linguistic Tools for Retrieval Systems. *Journal of the Association for Computing Machinery* 25: 52–66.

(38) Azar, Moche. 1970. Analyse morphologique automatique du texte hébreu de la Bible. *Tech. Rep. 12 et 19*, Faculte des Lettres et des Sciences Humaines, Nancy.

(39) Bashkansky, Guy and Uzzi Ornan. 1998. Monolingual translator workstation. In: *MT and the Information Soup: Proceedings of AMTA'98*. Springer. 136–149.

(40) Beesley, Ken. 1996. Arabic finite-state morphological analysis and generation. In: *Proceedings of COLING-96, the 16th International Conference on Computational Linguistics*.

(41) Beesley, Kenneth R. 1998. Arabic morphology using only finite-state operations. In: Michael Rosner (ed.). *Proceedings of the Workshop on Computational Approaches to Semitic languages*. COLING-ACL'98. 50–57.

(42) Beesley, Kenneth R. and Lauri Karttunen. Forthcoming. *Finite-State Morphology: Xerox Tools and Techniques*.

(43) Carmel, David and Yoelle Maarek. 1999. Morphological Disambiguation for Hebrew Search Systems. In: *Proceedings of te 4th International Workshop* NGITS-99, Israel. Lecture Notes in Computer Science 1649. Springer: 312–325.

(44) Choueka, Yaacov. 1980. Computerized Full-Text Retrieval Systems and Research in the Humanities: The Responsa Project. *Computers and the Humanities* 14: 153–169.

(45) Choueka, Yaacov and Serge Lusignan. 1985. Disambiguation by Short Context. *Computers and the Humanities* 19: 147–157.

(46) Dagan, Ido and Alon Itai. 1994. Word sense disambiguation using a second language monolingual corpus. *Computational Linguistics* 20: 563–596.

(47) Dahan Netzer, Yael. 1997. *HUGG — Unification-Based Grammar for the Generation of Hebrew Noun Phrases*. Master's thesis, Ben-Gurion University of the Negev, Department of Computer Science, Faculty of Natural Sciences, Be'er Sheva, Israel.

(48) Dahan Netzer, Yael and Michael Elhadad. 1998. Generating determiners and quantifiers in Hebrew. In: Michael Rosner (ed.). *Proceedings of the Workshop on Computational Approaches to Semitic Languages* (COLING/ACL'98). 82–88.

(49) Dahan Netzer, Yael and Michael Elhadad. 1998. Generation of noun compounds in Hebrew: Can syntactic knowledge be fully encapsulated? In: Eduard Hovy (ed.). *Proceedings of the Ninth International Workshop on Natural Language Generation.* Association for Computational Linguistics. 168–177.

(50) Dahan Netzer, Yael and Michael Elhadad. 1999. Hebrew-English Generation of Possessives and Partitives: Raising the Input Abstraction Level. In: *Proceedings of the 37th meeting of the ACL.* 144–151.

(51) Dalrymple, Mary, Ronald M. Kaplan, John T. Maxwell and Annie Zaenen (eds.). 1995. *Formal Issues in Lexical-Functional Grammar, vol. 47 of CSLI lecture notes.* CSLI, Stanford, CA.

(52) Fraenkel, Aviezri S. 1976. All about the Responsa Retrieval Project — what you always wanted to know but were afraid to ask. *Jurimetrics Journal* 16: 149–156.

(53) Glinert, Lewis. 1989. *The Grammar of Modern Hebrew.* Cambridge: Cambridge University Press.

(54) Haddock, Nicholas, Ewan Klein and Glyn Morill (eds.). 1987. *Categorial Grammar, Unification and Parsing*, vol. 1 of *Working Papers in Cognitive Science.* University of Edinburgh, Center for Cognitive Science.

(55) Herz, J. and M. Rimon. 1991. Local Syntactic Constraints. In: *Proceedings of the Second International Workshop on Parsing Technologies.*

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

(57) Joshi, A. K. 1987. An Introduction to Tree Adjoining Grammars. In: A. Manaster-Ramer (ed.). *Mathematics of Language.* Amsterdam: John Benjamins.

(58) Kaplan, Ronald and Joan Bresnan. 1982. Lexical functional grammar: A

formal system for Grammatical Representation. In: J. Bresnan (ed.). *The Mental Representation of Grammatical Relations*. Cambridge, Mass.: MIT Press. 173–281.

(59) Kaplan, Ronald M. and Martin Kay. 1994. Regular Models of Phonological Rule Systems. *Computational Linguistics* 20: 331–378.

(60) Karttunen, Lauri, Jean-Pierre Chanod, Gregory Grefenstette and Anne Schiller. 1996. Regular Expressions for Language Engineering. *Natural Language Engineering* 2/4: 305–328.

(61) Kiraz, George Anton. 2000. Multitiered Nonlinear Morphology Using Multitape Finite Automata: a Case Study on Syriac and Arabic. *Computational Linguistics* 26: 77–105.

(62) Koskenniemi, Kimmo. 1983. *Two-Level Morphology: a General Computational Model for Word-Form Recognition and Production*. The Department of General Linguistics, University of Helsinki.

(63) Lavie, Alon, Alon Itai, Uzzi Ornan and Mori Rimon. 1988. On the Applicability of Two-Level Morphology to the Inflection of Hebrew Verbs. *Tech. Rep.* 513, Department of Computer Science, Technion.

(64) Lavie, Alon, Alon Itai, Uzzi Ornan and Mori Rimon. 1988. On the Applicability of Two-Level Morphology to the Inflection of Hebrew Verbs. In: *Proceedings of the International Conference of the ALLC*.

(65) Lazewnik, Rabbi Grainom. 1970. Construction of an Algorithm for Stem Recognition in the Hebrew Language. *Hebrew Computational Linguistics* 2: 84–101.

(66) Levinger, Moshe, Uzzi Ornan and Alon Itai. 1995. Learning morpho-lexical probabilities from an untagged corpus with an application to Hebrew. *Computational Linguistics* 21/3: 383–404.

(67) Mani, Anderjeet. 2001. *Automatic Summarization*. Amsterdam: Johns Benjamins.

(68) Mani, Anderjeet and Mark T. Maybury (eds.). 1999. *Advances in Automatic Text Summarization*. Cambridge, Mass.: MIT Press.

(69) Mohri, Mehryar. 1996. On Some Applications of Finite-State Automata Theory to Natural Language Processing. *Natural Language Engineering* 2: 61–80.

(70) Mohri, Mehryar, Fernando Pereira and Michael Riley. 1998. A rational design for a Weighted Finite-State Transducer Library. In: *Lecture Notes in Computer Science*. Springer. No. 1436.

(71) Morgenbrod, M. and E. Serifi. 1976. Computer-Analysed Aspects of Hebrew Verbs. *Hebrew Computational Linguistics* 10, E1-17.

(72) Morgenbrod, M. and E. Serifi. 1977. Computer-Analysed Aspects of Hebrew Verbs: Mathematical Models. *Hebrew Computational Linguistics* 12, E1-18.

(73) Morgenbrod, M. and E. Serifi. 1978. Computer-Analysed Aspects of Hebrew Verbs: The Binjanim Structure. *Hebrew computational linguistics* 14, V–XV.

(74) Nirenburg, Sergei and Yosef Ben-Asher. 1984. HUHU — the Hebrew University Hebrew Understander. *Computer Languages 9*, 3/4.

(75) Ornan, Uzzi. 1986. Phonemic Script: A Central Vehicle for Processing Natural Language — the Case of Hebrew. *Tech. Rep.* 88.181. Haifa: IBM Research Center.

(76) Ornan, Uzzi. 1994. Basic Concepts in "Romanization" of Scripts. *Tech. Rep. LCL* 94–95. Haifa: Laboratory for Computational Linguistics, Technion.

(77) Ornan, Uzzi and Israel Gutter. 2000. Machine Translation by Semantic Features. In: Derek Lewis and Ruslan Mitkov (eds.). *Machine Translation and Multilingual Applications in the New Millennium.*

(78) Ornan, Uzzi and Katz, Michael. 1995. A New Program for Hebrew Index Based on the Phonemic Script. *Tech. Rep. LCL* 94-7, Laboratory for Computational Linguistics, Technion, Haifa, Israel.

(79) Pollard, Carl and Sag, Ivan A. 1987. *Information Based Syntax and Semantics*. No. 13 in CSLI Lecture Notes. CSLI.

(80) Pollard, Carl and Ivan A. Sag. 1994. *Head-Driven Phrase Structure Grammar*. University of Chicago Press and CSLI Publications.

(81) Price, James D. 1969. An Algorithm for Generating Hebrew Words. *Hebrew Computational Linguistics 1*, 51–54. Reprinted from *Computer Studies in the Humanities and Verbal Behavior*, 1:84–102, 1969.

(82) Price, James D. 1970. The Development of a Theoretical Basis for Machine Aids for Translation from Hebrew to English. *Hebrew Computational Linguistics* 2, 65–83. Abstract of a Doctoral Dissertation, The Dropsie College for Hebrew and Cognate Learning, Philadelphia.

(83) Price, James D. 1971. An Algorithm for Analyzing Hebrew Words. *Computer Studies in the Humanities and Verbal Behavior* 3/2: 137–165.

(84) Price, James D. 1971. A Computerized Phrase Structure Grammar (Modern Hebrew). Franklin Institute Report F-C2585-1/2/3/4.

(85) Roche, Emmanuel and Yves Schabes (eds.). 1997. *Finite-State Language Processing*. Language, Speech and Communication. Cambridge: MIT Press.

(86) Samuelsdorff, Paul Otto. 1980. Computational Analysis of Modern Hebrew. *Hebrew Computational Linguistics* 16, IV–XVI.

(87) Segal, Erel. 1997. *Morphological analyzer for unvocalized hebrew words.* available from http://www.cs.technion.ac.il/erelsgl/hmntx.zip.

(88) Shieber, Stuart M. 1986. *An Introduction to Unification Based Approaches to Grammar*. CSLI Lecture Notes. CSLI.

(89) Shieber, Stuart M., Hans Uszkoreit, Fernando C. N. Pereira, J. J. Robinson and M. Tyson. 1983. The Formalism and Implementation of PATR-II. In: *Research on Interactive Acquisition and Use of Knowledge*. Menlo Park, Cal.: SRI International.

(90) Sima'an, Khalil, Alon Itai, Yoad Winter, Alon Altman and N. Nativ. To appear. Building a tree-bank of Modern Hebrew text. *Traitment Automatique des Langues*.

(91) Skoblikov, Victoria. 2000. *Feature-based Computational Lexicon of Hebrew Verbs*. Master's thesis, Technion, Israel Institute of Technology, Haifa, Israel.

(92) Sproat, R. W. 1992. *Morphology and Computation*. Cambridge: MIT Press.

(93) Steedman, Mark. 2000. *The Syntactic Process*. Language, Speech and Communication. Cambridge: The MIT Press.

(94) Talmon, Rafi and Shuly Wintner. 2001. Computational Processing of Spoken North Israeli Arabic. In: Arabic Language Processing: Status and Prospects, Association for Computational Linguistics. 124–126.

(95) Vaillette, Nathan. 2001. Hebrew Relative Clauses in HPSG. In: Dan Flickinger and Andreas Kathol (eds.). *Proceedings of the 7th International Conference on Head-Driven Phrase Structure Grammar*. CSLI Publications.

(96) van der Toorn, A. J. 1971. Automatic Reading of Handwritten Hebrew. *Hebrew Computational Linguistics* 4: 83–99.

(97) van Noord, Gertjan and Dale Gerdemann. 2001. Finite State Transducers with Predicates and Identity. *Grammars* 4/3.

(98) Wintner, Shuly. 1991. *Syntactic Analysis of Hebrew Sentences*. Master's thesis, Technion, Israel Institute of Technology, Haifa, Israel. In Hebrew, abstract in English.

(99) Wintner, Shuly. 1997. *An Abstract Machine for Unification Grammars*. PhD thesis, Technion — Israel Institute of Technology, Haifa, Israel.

(100) Wintner, Shuly. 1998. Towards a Linguistically Motivated Computational Grammar for Hebrew. In: Michael Rosner (ed.). *Proceedings of the Workshop on Computational Approaches to Semitic Languages*. COLING-ACL'98. Association for Computational Linguistics. 82–88.

(101) Wintner, Shuly (ed.). 2001. Israeli Seminar on Computational Linguistics (ISCOL'01).

(102) Wintner, Shuly and Uzzi Ornan. 1991. Computational Models for Syntactic Analysis — their fitness for writing a computational grammar for Hebrew. In: *Proceedings of the Bar-Ilan Symposium on Foundations of Artificial Intelligence*. Also as CIS Report 9103, Center for Intelligent Systems, Technion, May 1991.

(103) Wintner, Shuly and Uzzi Ornan. 1991. Syntactic Analysis of Hebrew Sentences. In: *Proceedings of the 8th Israeli Symposium on Artificial Intelligence and Computer Vision*. Information Processing Association of Israel. 201–230.

(104) Wintner, Shuly and Uzzi Ornan. 1996. Syntactic Analysis of Hebrew Sentences. *Natural Language Engineering* 1: 261–288.