המורכבות בפיתוח מנועי חיפוש בעברית

יעקב שויקה

״המורכבות בפיתוח מנועי חיפוש בעברית״

ראשי פרקים להרצאתו של

פרופ׳ יעקב שויקה

שניתנה ב 17/12/1997

בקמפוס גבעת רם, בפני פורום מנהלי אתרי אינטרנט ממשלתיים

רשם: דני גולן , שימר: עופר דרורי

השפה האנגלית היא ממשפחת השפות ההודו-אירופיות, מאפיין השפה - הרבה צורות יסוד, מעט הטיות וצורות משנה.

השפה עברית היא ממשפחת השפות השמיות, מאפיין השפה - מעט צורות יסוד אבל עושר הטיות וצורות משנה.

קבוצה	Group	כמות באנגלית	כמות בעברית	הסבר
שורשים	STEM	40,000	5,000	שורש המילה
ערכים	LEMMAS	150,000	35,000	צורת יסוד למילה (כפי שמופיעה במילון) כולל מילים לועזיות שהשתלבו בשפה.
צורות	WORDS	1,000,000	70,000,000	כל המילים כולל הטיות. כל המחרוזות של מילים תקינות.

מסקנה: רמת הסיבוכיות של המורפולוגיה בעברית גבוהה יותר מאשר באנגלית, בגלל ריבוי הצורות.

- מכל **שם תואר** (וגם מהרבה **שמות עצם**) במילון ניתן לפתח כמה מאות (ולפעמים עד 1200) צורות, ע״י:

מין: זכר, נקבה

מספר: יחיד, זוגי, רבים

אותיות שימוש: מ.ש.ה.ו.כ.ל.ב. וצירופיהם

כינוי: גוף שלי, שלך, שלהן, ...

כתיב: מלא/חסר (י,ו ,א), כיסא/כסא, ברכותיי/ברכותי,...

כתיב שונה: א/ה, ז/ס, ס/ש, ... לדוגמה: טבלא/ טבלה, מוסיקה/ מוזיקה, ...

- מפועל ניתן לפתח אלפי צורות

7 בניינים: פעל, נפעל, הפעיל, ...

4 אופנים: עבר, הווה, עתיד, ציווי

12 גופים: אני, אתה, ...

אותיות שימוש: מ.ש.ה.ו.כ.ל.ב. וצירופיהם

כינוי פעול: ראיתיך, ...

כתיב מלא: תישמרי, אשמור

דוגמאות למילים שקשה לשחזר את מקורן המילוני:

וכש פי ותי הם רק "פ" נותרה מתוך המקור המילוני –"פה"

כש ת ראו ני רק שתי אותיות משותפות למקור המילוני "ראה"

ל כ ש י כ ו ה ו השורש "נכה", המקור המילוני (צורת העבר) "הכה". גם כאן נשארה רק כ מתוך המקור.

ברור אם כך שלצורכי אחזור מידע נדרש פיתוח תוכנת שיחזור מורכבת כדי לאתר את המקור המילוני של מילה.

אם מחפשים לדוגמא במאגר את הביטוי "הפלת מטוסים", הרי שהרבה מחרוזות (חלופות מורפולוגיות של המלים שבביטוי) אמורות להשתתף בחיפוש שכזה:

הפלת, כשההפלות, הפלנו, ... טייסים, טיסות, מטוסנו, ...

אפשר אמנם לבצע חיפוש עם אופרטורים להרחבה פורמאלית כגון:

ילד, *תאונה , מטוס*, ...

אבל זה לא מספק.

במקרה, לדוגמה, של "לקיחת שוחד", נוכל לכאורה להסתפק בחיפוש עם כוכביות למילה *לקח*, אבל חיפוש כזה לא ימצא את "א**קח**" "י"**קחו**" "**לקיח**ת" "**לוקח**" שהן מילים שאמורות להשתתף בחיפוש, ולעומת זאת נקבל מילים כמו "מת**לקח**" "מ**לקח**יים" שאינן שייכות לחיפוש אבל עונות על התנאים.

דוגמא אחרת, בחיפוש על אלימות נגד נשים באמצעות הביטוי *הכאה *אישה

לא נקבל את הצורות "מכות" "הכה" "מוכות" "יכו" "להכות" שהן ללא ספק רלוונטיות

ובחיפוש על "בית" לפי *בית נקבל "ח**בית**ה" "מג**בית**" "**בית**ר" מילים שבוודאי אינן קשורות לבית.

ולעומת זאת לא נקבל "בתים" "בתי" "שבתיהם" שכולן רלוונטיות.

מסקנה: מנוע חיפוש פשוט ללא לוגיקה לשונית אינו קביל במאגרים בהם חשוב למצות את החומר השייך לנושא המבוקש, כמו מאגרים משפטיים, אקדמיים, מדעיים, צבאיים וכדומה.

בניסוי שנעשה על מאגר מכון סאלד המכיל כ 300 תקצירי מאמרים בענייני חינוך, ומטרתו היתה לאתר חומר המכיל "הצלחה" או "לימוד", נעשה החיפוש פעם על המחרוזת עצמה, פעם על המחרוזת עם הרחבות פורמאליות, ופעם עם הרחבות מורפולוגיות, והתוצאות היו:

סוג חיפוש	מאמרים	מופעים של המילה	הסבר
הצלחה	5	7	חיפוש מדויק
*הצלחה	13	14	הרחבה פורמאלית
הצלחה	18	19	חיפוש מורפולוגי מלא
לימוד	9	9	חיפוש מדויק
לימוד	97	119	הרחבה פורמאלית
לימוד	189	267	חיפוש מורפולוגי מלא

קיימות שתי גישות לטפל בבעיה:

<u>גישת הסינתזה</u>:

פיתוח כל החלופות המורפולוגיות הקשורות למילת החיפוש, וביצוע החיפוש על כל החלופות שפותחו.

הבעיה בשיטה זו היא שתהליך הפיתוח און ליין גוזל הרבה זמן ביצוע וכן מתבצע חיפוש מיותר על צורות שאינן בטקסט.

<u>גישת האנליזה</u>:

ניתוח דקדוקי מורפולוגי של כל מילה במאגר בזמן האינדוקס והצמדת "ערך יסוד" לכל מילה.
כאשר מבצעים חיפוש על מילה, מתבצע ניתוח ופירוק מורפולוגי של המילה כדי לקבל את ערך היסוד שלה, החיפוש מתבצע לפי ערך היסוד ויזוהו כל המילים שבמאגר עם אותו ערך יסוד.

<u>מערכת "מילים" של שויקה</u>

מערכת למיכון ידע מורפולוגי בעברית בת זמננו, פותחה בתשמ"ט במסגרת מט"ח – המרכז לטכנולוגיה חינוכית

מאפיינים:

מקיפה מצד אחד ומדויקת מצד שני; כל מילה בעברית בת זמננו תקבל את כל הניתוחים המורפולוגיים הנכונים.

טיפול בכתיב מלא וחסר כאחד

כתובה בשפת C, מערכות הפעלה PC וגם VAX

תגובה מהירה מאוד, מאות ניתוחים בשנייה

זיכרון נדרש: חצי מגה בס"ה

טיפול בראשי תיבות, בקיצורים, במחרוזות מספריות ובשמות פרטיים של ישויות (מקומות, מטבעות וכד')

כנגזרת ממערכת "מילים", פותח גם:

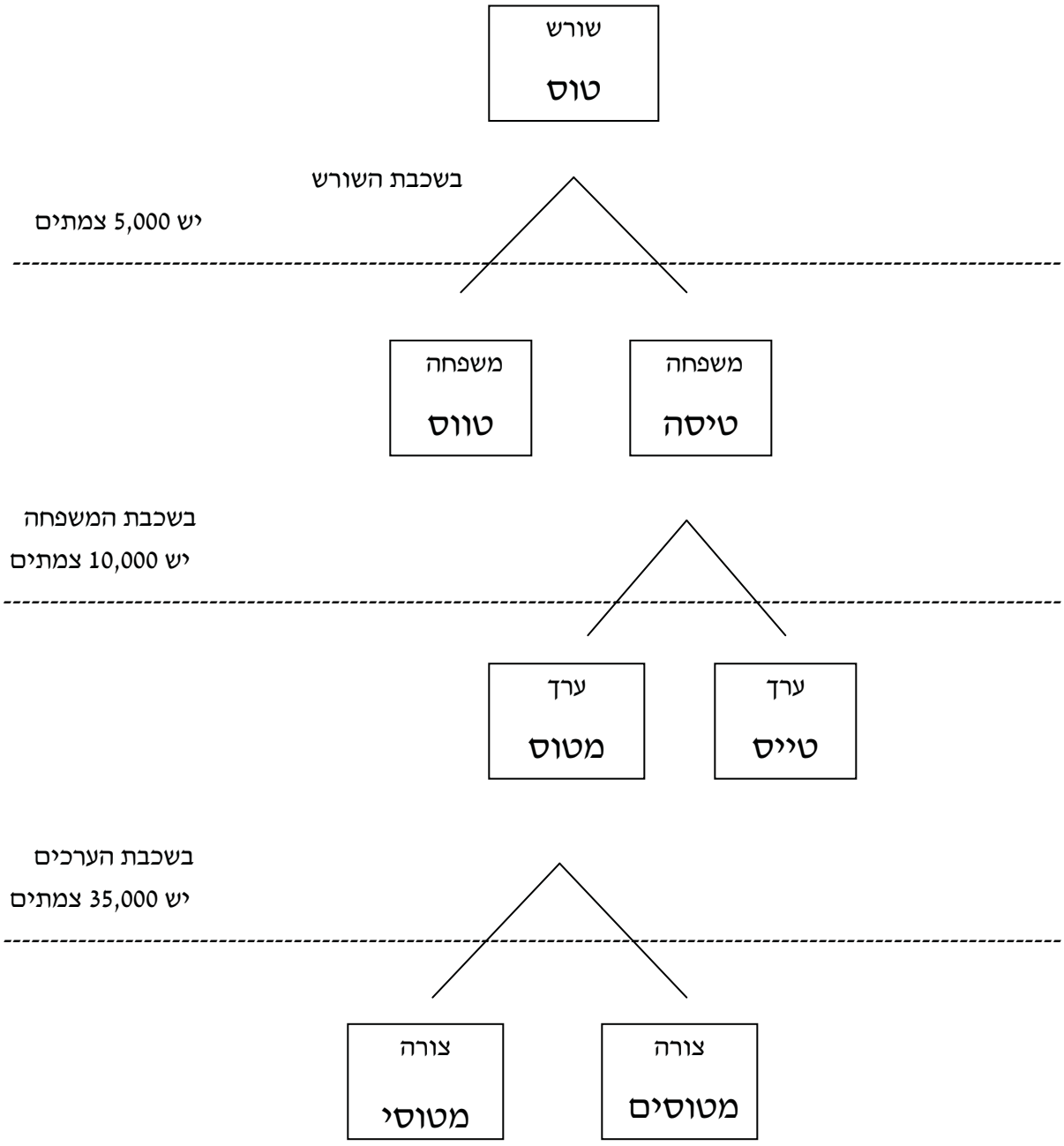
"נקדן" - מערכת המנקדת את הטקסט לפי כללי הניקוד העברי, כולל פענוח עמימות לפי ההקשר

היסטוריה של מערכות מורפולוגיות שפותחו ע"י שויקה:

- בשנת 1963 פיתח במודיעין מערכת לניתוח מורפולוגי מלא לעברית בשפת אסמבלר של מחשב פילקו בממר"מ, ועליה נבנתה תוכנה לבניית קונקורדנציות ממוינות לפי שורשים וערכים, והיא הופעלה על 80 העמודים הראשונים של "ימי צקלג" לס. יזהר. המערכת תוארה במאמר ב"לשוננו", ביטאון האקדמיה ללשון העברית ב 1964.
- בשנים 1968 - 1970 פיתח פתרון ביניים של סינתזה לשילוב מנוע מורפולוגי במערכת השו"ת – מערכת לחיפוש בטקסט מלא במאגרים רבניים גדולים
- בסוף שנות השבעים פיתח מערכת משולבת סנטזה/אנליזה לשילוב בגרסה המקוונת (און-ליין) של מערכת השו"ת (הכוללת גם טיפול ראשוני בארמית), מנוע שממשיך להיות משולב, עד היום, כמות שהוא, בתקליטורים של השו"ת.
- בשנת 1989 פיתח את מערכת מלי"ם, המשולבת כיום במספר ניכר של מערכות אחזור מידע בטקסט מלא בעברית; ממנה גם נגזר בודק כתיב שהוא היום בודק הכתיב הסטנדרטי המשולב במערכת WORD בעברית של מיקרוסופט

לאחרונה פותחו מספר מנועים מורפולוגיים חלקיים מאוד בעברית, שרובן גם אינן פעילות כיום בפועל, וביניהן מערכת שפותחה במעבדות י.ב.מ. בחיפה.

עץ השפה



הוספת המשפחה מרחיבה את הטבלא שבראשית ההרצאה:
משפחה מכילה את כל הערכים בעלי אותו שורש ואותו שדה סמנטי.

חיפוש לפי שורש אינו יעיל ואינו מומלץ, שהרי בשורש אחד עשויים להיות ערכים משדות סמנטיים שונים; כך למשל:

- השורש ס.פ.ר. הכולל לפחות שלוש משפחות שונות: ספר שקוראים, לספור מספרים, מספריים לגזירה.
- השורש טוס: כולל לפחות שתי משפחות: מטוס, טווס
- השורש 'לקח' כולל לפחות שתי משפחות: לקיחה של דבר, התלקח באש.
- השורש 'להב' כולל לפחות שלוש משפחות: התלהב מרעיון, להבה של אש, להב של סכין.

רצוי על כן לבצע חיפוש שירוץ על כל הערכים שהם מאותה משפחה כמו מלת החיפוש.

שכבה	כמות צמתים
שורשים	5,000
משפחות	10,000
ערכים	35,000
צורות	70,000,000

חמש רמות בהרחבה: מחרוזת מדויקת, הרחבה צורנית, ערך, משפחה, תזאורוס

בעיית המילים הנפוצות STOP WORDS

באנגלית הבעיה יחסית פשוטה: יש 100 מילים נפוצות שאין צורך למפתח אותן והן מקיפות כ - 40% מהטקסט. למילים הנפוצות (כמו and, or, the, on, of) אין כפל משמעות.

בעברית הבעיה מורכבת יותר מכמה טעמים:

יש כ 600 מילים נפוצות בצורותיהן השונות. הן מכסות בממוצע רק כ 20% מהטקסט.

לחלק מהמילים יש כפל משמעות (כך שלא ניתן לוותר עליהן), כגון "את", שהיא גם "את חפירה"; "אבל" שיש לה גם משמעות של "אבל לאומי", ו "אם" שהיא גם "אם יצחק".

בעיית עמימות Ambiguity

התופעה קיימת בכל השפות הטבעיות. כך, למשל, באנגלית:

BANK = בנק, גדה.

RECORD = תקליט, רשומה, שיא.

בעברית הבעיה מוכפלת, שהלוא בעברית ללא ניקוד חסרות כל התנועות (VOWELS)

למשפט "הרכבת הטיל תהיה דבר מסוכן", לדוגמה, יש 480 אפשרויות לקרוא אחת ורק אחת מהן נכונה:

המשפט המקורי	הרכבת	הטיל	תהיה	דבר	מסוכן
צורות נוספות לקרוא	רכבת ישראל הרכבת על סוס	טייל בשביל הטיל מימיו		מחלת דבר דבר השבוע	מסוכן חשאי

דוגמא נוספת "הציר קודח והקרן מנגנת"

ציר = דלת, כאב, שגריר.

קודח = מחום, עם מקדחה.

קרן = כלי נגינה, קרן חייה.