– Hspell

בודק איות (ומנתח מורפולוגי)

עברי חופשי

דן קניגסברג מדעי המחשב, הטכניון

נדב הראל מעבדת המחקר חיפה, יבמ

danken@cs
nyh@math .technion.ac.il

מיזם Hspell

הצורך:

- תכנות חופשיות מרכזיות, כמו מעבדי תמלילים, תומכות בעברית.
- לא היה בודק איות עברי חופשי.
- לא הייתה אפילו רשימת מילים עבריות חופשית.

הפתרון:

- כתיבה עצמית הדרגתית של בודק איות.
- גרסה 0.1 של Hspell יצאה בדצמבר 2002, גרסה 0.9 תצא בקרוב.

עקרונות הפיתוח

- פיתוח ב"חדר נקי": אין העתקה של רשימות מילים.
- שאיפות מעשיות בשלב ראשון: בדיקת איות לכתיב חסר־ניקוד.

  בחרנו בתקן **הכתיב חסר הניקוד** של האקדמיה ללשון העברית (1949–1993), אשר מופיע במלואו כנספח ברוב המילונים.
- אפשרות לשימושים מתקדמים יותר בעתיד הקרוב והרחוק.

  מנתח צורני קיים מגרסה 0.7.
- תוצרי הפיתוח יהיו זמינים לכול, כמיטב המסורת של המחקר המדעי והתכנה החופשית.

מבנה הפתרון

הגישה שלנו – סינתטית – כוללת שלושה שלבים:

1. **איסוף** (ידני) של מילות בסיס.
2. **הטיה** (perl): בנייה חצי־אוטומטית של כל הנטיות של מילות הבסיס.
3. **בדיקת איות** (C): בהינתן טקסט עברי, עוברים מילה־מילה ובודקים אם היא נמצאת ברשימת הנטיות החוקיות.
   מתיר את אותיות השימוש **מש"ה וכל"ב** בהתאם לסוג המילה.

יתרונות וחסרונות

לשיטת הסינתזה יתרונות רבים יחסית לגזירה־לאחור:

- קל לשלב בבודקי איות רב־לשוניים, המבוססים על רשימות מילים
- הפרדה לשלבים פשוטים יותר ובלתי תלויים
- הפרדה קלה בין המפתחים
- אלגוריתם זמן־ריצה פשוט ומהיר
- קוד פשוט וקריא יותר

חסרונות:

- צריכת זיכרון (ללא suffix compression).
- משך אתחול ארוך יותר (לא בעיה אמתית, שבריר שנייה).

1. איסוף

עד עתה (גרסה 0.9α) אספנו

- 10,712 **שמות עצם**
- 2,996 **תארים**
- 5,205 **פעלים**
- 2,127 **מילים אחרות**

```
כלב ע
ירוק ת
שמר פ,קל_אפעול+,נפ,פי+,פו,הת
בפרוטרוט
```

אבל גם

```
צומת ע,ים,אבד_ו
עוף ע,ות
עיפרון ע,ות,אבד_י
חסר פ,קל_אפעל,בינוני_שמן,הפ,הו,אין_שם_פעולה,אין_פעול
```

2. הטיה

- **שמות עצם**. הטיה אוטומטית ליחיד, רבים, נפרד, נסמך, כינויים.

  כלב כלב־ כלבי כלבנו כלבך כלבך כלבכם כלבכן כלבו כלבה כלבן כלבם

  כלבים כלבי־ כלביי כלבינו כלביך כלבייך כלביכם כלביכן כלביו כלביה כלביהן כלביהם

- **תארים**. הטיה אוטומטית ליחיד, רבים, נסמך, זכר, נקבה.

  ירוק ירוק־ ירוקים ירוקי־

  ירוקה ירוקת־ ירוקות ירוקות־

- **פעלים**. הטיה אוטומטית לבניינים אפשריים, זמנים, שמות פעולה וכו'.

לשמור שמירה	שמרתי שמרת שמרת שמר שמרה שמרנו שמרתם שמרתן שמרו	שומר שומרת שומרים שומרות שמור שמורה שמורת־ שמורים שמורי־ שמורות	אשמור תשמור תשמרי ישמור תשמור נשמור תשמרו תשמורנה ישמרו	שמור שמרי שמרו שמורנה	לשמרני וכו' שמרתיו וכו'

2. הטיה (המשך)

- **מילים נוספות**, ללא הטיה אוטומטית:
  - מילות ונטיותיהן (של, שלי, ...)
  - תוארי הפועל (היכן, שלשום, הרבה, ...)
  - שמות גוף (אני, אלה, מיהו, ...)
  - מקצועות, תחביבים, אמונות, סגנונות (ביולוגיה, כדורגל, קומוניזם, ...)
  - שמות עבריים
  - ועוד ועוד

ובסך הכל 444,403 נטיות ומילים.

רשימת מילים זו תופסת מקום זניח על הדיסק: 147,379 בתים. (כולל מידע על תחיליות חוקיות לכל מילה, ללא מנתח צורני מלא).

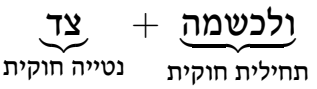
הרשימה זמינה באופן חופשי לכל דורש (תחת רישיון GPL) וחלקה שולב בידי שלמה יונה במאגר של מיל"ה.

3. בדיקת איות

רשימת הנטיות נטענת לזיכרון, ומאוחסנת במבנה נתונים יעיל.

בהינתן טקסט עברי הוא מפורק למילים.

מנסים לפרק כל מילה לתחילית חוקית + נטייה מרשימת הנטיות



אם לא הצלחנו למצוא פירוק – זו שגיאת כתיב.

תחילות

מצאנו (בסיוע שלמה יונה) דקדוק חסר־הקשר אשר יוצר את כל התחיליות בעברית:

תחילית → ו ש כש בכלמה | מבל | בכלמש

ו → ו | ε

ש → ש | ה$_{\text{השאלה}}$ | ε

כש → כש | מש | לכש | ε

בכלמה → ה$_{\text{הידיעה}}$

| כ$_1$ (ב | ל | מ | ε) כ$_2$

| כ$_1$ (ב | ל | מה)

כ$_1$ → כ$_{\text{"כמו"}}$ | ε

כ$_2$ → כ$_{\text{כמת}}$ | ε

מבל → מ (ב | ל)

בכלמש → (ב | כ | ל | מ) ש

(יוצר גם תחיליות אקזוטיות מאוד "כמכאלף ארובות נשפך עליי הגשם", כ־360 תחיליות בסך הכול)

התאמת תחילית־מילה

ייצרנו את כל הנטיות החוקיות ואת כל התחיליות החוקיות – אבל לא כל שילוב שלהן חוקי!

וכש+קפוצנה, ולכשמה+שמרתי

פתרנו את הבעיה כדלקמן:

- כל תחילית **מספקת** מספר *סיביות* ע"פ האותיות שמרכיבות אותה.
- כל נטייה **דורשת** *סיבית* אחת, בהתאם לגזירה שלה.

  לעתים, כמה נטיות שונות נראות על פני השטח כנטייה אחת. במקרה זה, הנטייה תדרוש כמה סיביות.
- מותר לצרף תחילית לנטייה רק אם היא מספקת לפחות אחת מהסיביות הנדרשות.

לדוגמה, המילה "קפוצנה" דורשת את הסיבית IMPER, שהתחילית "ו־" מספקת אך "וכש־" אינה מספקת. לכן השילוב ו+קפוצנה חוקי בעוד וכש+קפוצנה איננו.
המילה "שמרתי" דורשת את הסיבית NONDEF ולכן אינה יכולה להשתלב עם התחילית "ולכשמה־" אשר איננה מספקת סיבית זו. השילוב וכש+שמרתי חוקי כי התחילית "וכש־" מספקת את כל הסיביות ש"שמרתי" דורשת.

ניתוח צורני

- בזכות הגישה הסינתטית, אנו מקבלים כמעט "במתנה" ניתוח צורני.
- בשלב ההטיה, זוכרים מהי מילת הבסיס של כל נטייה, וכיצד היא הופקה.

```
-----
שמר פ,עבר,הוא
שמרנ פ,עבר,הוא,כינוי/אנ
שמרך פ,עבר,הוא,כינוי/אתה
...
```

- בשלב האיות טוענים את המידע לזיכרון, במבנה נתונים לא יעיל במיוחד.
- בהינתן מילה עברית:
  - נגזום מראשיתה תחילית
  - נבדוק אם השאר הוא נטייה חוקית
  - נבדוק אם נטייה זו משתלבת כראוי עם התחילית שלה.
  - אם השילוב חוקי, קל כעת להדפיס את מקורה של המילה, ולחפש גיזומים נוספים.

זה עובד!

מציאת שגיאות כתיב ותיקוניהן המומלצים (מנשק טקסטואלי)

57 גרדנר ←

32 פונקצית ← פונקציית, פונקציות

25 קדקד ← קדקוד

לפי כללי הכתיב חסר הניקוד ("כתיב מלא") של האקדמיה ללשון העברית, תנועת "o"
שמקורה בחולם תיכתב בוי"ו. אולם זה לא המצב במקרה של קמץ קטן או חטף-קמץ - אז
תיכתב וי"ו רק אם במילה הבסיסית היה חולם.
לפיכך במילים שנכתבות בקמץ קטן או בחטף-קמץ בכל צורות המילה, אין להוסיף וי"ו.
לדוגמה: "תכנית", "קרבן", "חכמה", "אנייה", "נכרי", וכן "קדקוד".
יש לציין שלמרות שכלל זה נתקבל ע"י האקדמיה, הוא אינו מקובל על מקצת הבלשנים,
שמעדיפים היו לסמן וי"ו בכל מקום בו נהגית תנועה "o", כולל הקמץ הקטן וחטף הקמץ.

10 מימדית ← ממדית

7 כרצונינו ← כרצוננו

3 הפונציה ←

3 בוקטור ← בווקטור

2 המדידיות ← המדידות

2 השיוויון ← השוויון

זה עובד! (המשך)

Hspell **מספק מנשקי תכנות** (APIs): `C`, `ispell`.

באמצעותם שולב (בידי מפתחים אחרים) עם:

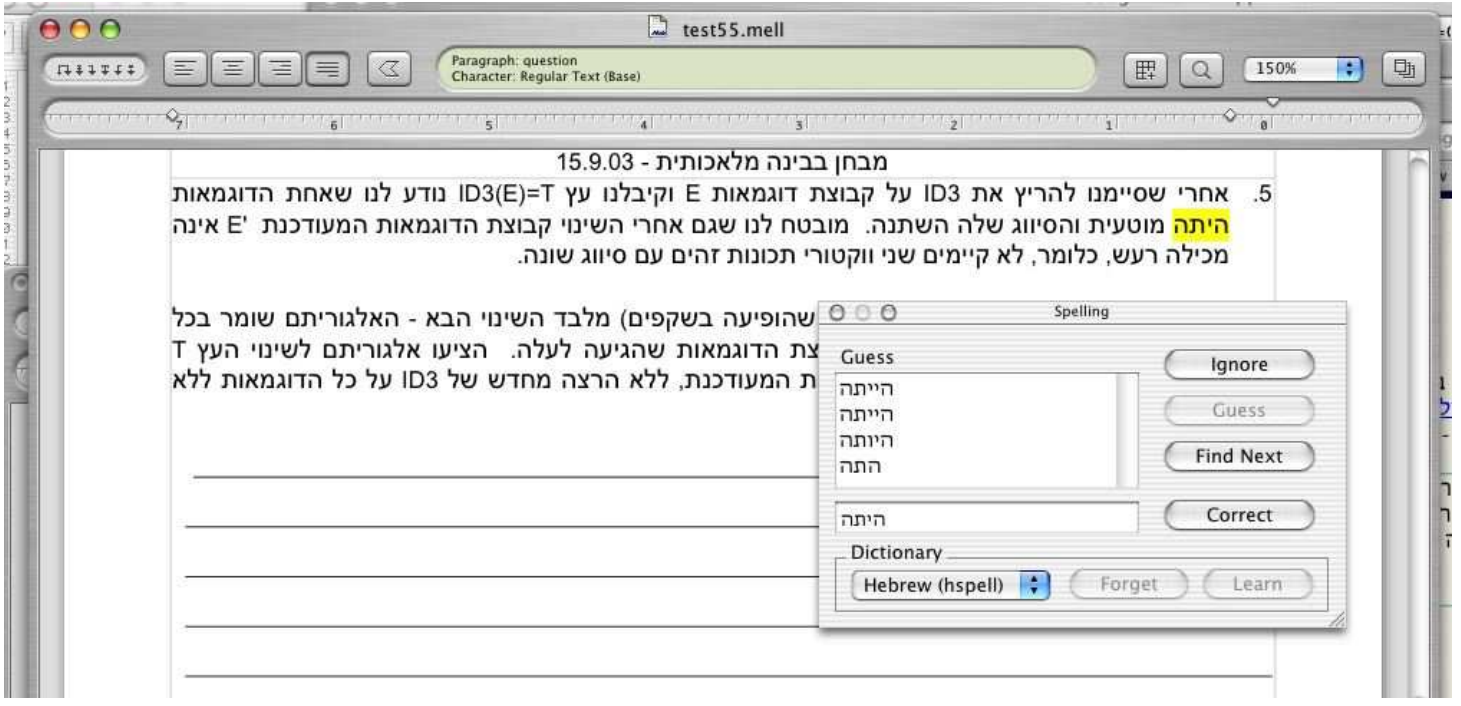
- **מעבדי תמלילים**: OpenOffice, Abiword, LyX, Mellel
- **שולחנות עבודה**: KDE, gnome
- **עורכי טקסט**: Emacs, Vim, Geresh
- **מערכות הפעלה**: Linux (Redhat, Debian, ...), FreeBSD, Mac OS X

לדוגמה, גרסה עברית של OpenOffice **באה עם** Hspell:


הספר עצמו, אשר בשפת המקור נקרא כ- "Anti-Anti: Tatsachen zur Judenfrage" (שהוא משחק מילים וקיצור לביטוי "אנטי-אנטישמיות") קובץ כאוסף דפים נפרדים שלא נכרכו, אשר קובצו לאגרון במעטפת קרטון קשיח, אולי כדי להקל על חלוקתם האפשרית במקרה הצורך. כמו גם להמשיך ולעדכנם כל פעם מחדש על ידי תוספות של דפים חדשים, אשר ניתן היה להזמין ללא כל בעייה במשרדי ההוצאה. ריפרנף ראשוני ב-781 העמודים (אשר חלקם גם עמודים כפולים) מראה כי עורכי הספר לקחו פרוייקט זה ברצינות הראויה, והקדישו לו ממיטב זמנם ומרצם, למען העלאת הנימוקים באופן שווה ושקול עבור כל נפש העתידה לבוא ולעלעל בדפים השונים. הקריאה בספר זה אינה מחייבת סדר מסויים, וניתן למעשה לפתוח בכל דף אפשרי, המוקדש לנושא אחר. הדפים עצמם קיבלו מספור בהתאם לנושאים השונים בהם טיפלו, שסודרו בתוכן העניינים לפי סדר הא"ב, לנוחיותו של הקורא, כשהמונח "אנטישמיות" (Antisemitismus) תחת האות A פותח אגרון זה, ומסיימו באות Z תחת הערך הגרמני "יהדות מפוררת" (Zersetzendes Judentum).

זה עובד! (המשך)

שילוב Hspell במקינטוש (מעבד תמלילים Mellel):



זה עובד! (המשך)

מנתח צורני (מורפולוגי):

כלבים	הרכבת	משטרה
כלבים: כלב(ע,ז,רבים)	**הרכבת:** הרכיב(פ,נ,2,יחיד,עבר) הרכיב(פ,ז,2,יחיד,עבר) הרכבה(ע,נ,יחיד,סמיכות) **ה+רכבת:** רכבת(ע,נ,יחיד) **ה+רכבת:** (ה"א השאלה) רכב(פ,נ,2,יחיד,עבר) רכב(פ,נ,2,יחיד,עבר) רכבת(ע,נ,יחיד,סמיכות)	**משטרה:** משטרה(ע,נ,יחיד) משטר(ע,ז,יחיד,של/היא) **מ+שטרה:** שטר(ע,ז,יחיד,של/היא)

ניתן לבחון אותו באתר

http://wassist.cs.technion.ac.il/ ~ danken/cgi-bin/hspell.cgi

מחקר ופיתוח עתידי

- הפחתה בצריכת הזיכרון – דחיסת סופיות.
- תמיכה בתקני איות אחרים (לא רק אקדמיה).
- תמיכה בניקוד
- איות נבון יותר (תלוי־הקשר)
- חיפוש בטקסט
- תיקון סגנון
- קריאה בקול
- ניקוד אוטומטי
- תרגום אוטומטי
- ועוד חלומות ...

סיכום

- אספנו את רוב המילים בעברית המודרנית (למעלה מ־20,000 מילות בסיס).
  סוף־סוף העברית הצטרפה לקבוצת השפות שרשימת המילים בהן זמינה.
- פיתחנו מאיית עברי – יחיד במינו בעולם היוניקס.
- קיבלנו "במתנה" מנתח מורפולוגי.
- הכול זמין באופן חופשי ברישיון GNU General Public License.

http://ivrix.org.il/projects/spell-checker