Hspell
בודק איות עברי חופשי

http://hspell.ivrix.org.il/

נדב הראל
דן קניגסברג

מועדון הלינוקס החיפאי
1 פברואר, 2010

מהו Hspell

- בודק איות עברי.
- גם מנתח צורני.
- רישיון חופשי (GPL).
- אוצר מילים נרחב: 23,486 מילים (גרסה 1.1).
- מקפיד על תקן הכתיב חסר הניקוד של האקדמיה ללשון העברית.
- נפוץ: משמש בהפצות לינוקס, אופן-אופיס, פיירפוקס, Gmail, ועוד.
- גרסה 0.1: 12/02. גרסה 1.1: 12/09.

מטרות ההרצאה

- סיכום 7 שנים מאז הגרסה הראשונה של Hspell
  - כיצד הוא פותח
  - כיצד הוא פועל
  - איך שולב בהפצות ובאפליקציות
  - תקן האיות
  - **סקירת החלטות שתרמו להצלחת המיזם (=פיתוח מהיר, איכות, ושימוש נרחב).**
- מעבר לבדיקת איות – הווה ועתיד.

רקע היסטורי

- 1983: "תכנה חופשית" (ריצ'רד סטולמן)
- 1991: לינוקס, קרנל חופשי ראשון, משלים מערכת הפעלה חופשית ראשונה.
- המשך שנות ה 90: נוחות, שימושיות, שפות:
  - הפצות לינוקס
  - Unicode
  - UTF8
  - Mosaic, Netscape, Mozilla
  - KDE
  - GNOME
  - Star Office

ראשיתו של hspell

- ינואר 2000: תמיכה עברית מתוכננת בפרויקטים הגדולים (דפדפן, DE, מעבד תמלילים)
- 12 בינואר, 2000 – הצורך בבודק איות עלה לראשונה ב ivrix-discuss
- באותו יום: רשימה של 1400 מילים (נטויות) מהתזה שלי.
- 21 בינואר: נטיית ש"ע אוטומטית
- 25 בינואר: דן תורם נטיית פעלים אוטומטית

החלטות ראשונות – מה לא לעשות

- אין טעם להתחנן שחברה מסחרית תשחרר בודק איות או רשימת מילים.
- לא מוסרי להעתיק רשימת מילים.
- אין טעם באיסוף אוטומטי של מילים ממסמכים:
  - לעולם לא יהיה שלם.
  - לא יהיה עקבי.
  - ייאספו גם שגיאות כתיב ומילים לא מילוניות.
- להתמקד בבדיקת איות ללא ניקוד. לא להתפזר עם ניקוד, פירושים ותרגומים – ל"רב מילים" זה דרש מאה שנות אדם!

כיצד Hspell כמעט נכשל

- אחרי פרץ עבודה ראשוני, כמעט התייאשנו:
- דיונים ברשימת דיוור הראו פחות התלהבות ויותר ניסיונות לשכנענו לשאוב רשימת מילים מבודק איות קיים, או לבקש תרומה של קוד מוכן.
- הגענו למסקנה ש ispell אינו טוב מספיק לצרכינו.
- הפיתוח נעצר לשנתיים.
- עד שהפצרות שני אנשים שכנעו אותנו עד כמה המיזם חשוב, וחזרנו לעבוד עליו.
- מסקנה: גם הערכת המשתמשים, וגם עניין אישי - שניהם חשובים לתכנה חופשית!

התלבטות ראשונה

	רשימת מילים "טפשה" ("גישה סינטתית")	ניתוח בזמן ריצה (גישה "אנליטית")
שלב פיתוח	איסוף מילות בסיס ורמזי נטייה	איסוף מילות בסיס ורמזי נטייה
שלב קומפילציה	הטיה אוטומטית של מילים אלו (לא כולל תחיליות)	-
שלב ריצה (בדיקת איות)	בדיקת כל מילה ברשימת המילים החוקיות, אולי בתוספת תחיליות מש"ה וכל"ב בהתאם לסוג המילה	ניסיון לנחש את מילת הבסיס של המילה הנתונה, ואז בדיקה האם הנטייה הנ"ל באמת חוקית
יתרונות	**אפשרות שילוב בבודקי איות מבוססי-רשימות קיימים.** קוד פשוט יותר.	צריכת זיכרון ודיסק נמוכה, בקלות

התלבטות ראשונה (המשך)

- החלטנו: **הגישה הסינתטית**.
- יצירת רשימות מילים:
  - איסוף ידני של מילות בסיס ורמזי נטייה
  - הטיה אוטומטית בעזרת תכניות perl
- בדיקת איות:
  - Ispell לא תמך כראוי בהוספת תחיליות
  - גרסה 0.1: בודק איות בפרל, אטי וזולל זיכרון.
  - מגרסה 0.6 (אוגוסט 2003): בודק איות יעיל ב C.
  - מגרסה 0.8 (יוני 2004): **בדיקת איות גם עם aspell**.

התלבטות ראשונה (המשך)

- בדיקת איות היום:
  - פקודה hspell, וספרייה, יעילות ובמסורת Unix .
  - בכל זאת, כמעט אף אחד לא משתמש בהן:
    - יצרנו רשימות מילים עבור aspell, myspell, hunspell.
    - מילון עברי כזה נכלל ברוב הפצות הלינוקס. מנגנון myspell קיים גם ב Mac OS X Snow Leopard.
    - פלג-אין בודק איות עברי באופן-אופיס ופיירפוקס משתמש במילון עברי דרך hunspell או myspell.
    - כך גם (כנראה) עושה gmail.
- **גישת רשימת המילים: בחירה נכונה וחשובה**

השוואת ביצועים (בדיקת איות בלבד)

hunspell -d he_IL (double-affix-compressed since Hspell 1.1)	aspell -l he	hspell	
3457 K	8458 K	**153 K** (2.7 bits per word!)	שטח דיסק (מילון, התאמת תחיליות)
0.15 sec	0.01 sec (he.rws mapped)	0.03 sec	זמן טעינת מילון
87,268	436,000	**3,800,000**	מילים לשנייה (בדיקה לבד, ללא תיקון)
10552 K	780 K + 8396 K	5744 K	צריכת זיכרון pmap -d

דחיסת המילון

חתול חתולי חתולים חתולך כלב כלבי כלבים כלבך

461,000 מילים, 3500 KB.

- דחיסת gzip מוצאת קטעים משותפים. 979 KB.

חתול 0י 0ם 2ך כלב 0י 0ם 2ך

- מידול זה כבר מקטין ל 1017 KB.
- דחיסת gzip מקטינה ל 110 KB !

תקן האיות

- אנו בודקים איות במלל ללא ניקוד.
- אך מהו "כתיב מלא" נכון?
- נבחר תקן: **הכתיב חסר הניקוד** של האקדמיה ללשון העברית.
- התקן במלואו מופיע כנספח ברוב המילונים, תקציר באתר האקדמיה.
- רק לאחרונה החל להופיע הכתיב חסר הניקוד בערכים במילונים. האם מוזכר בבתי ספר?
- מסמך שלנו "סוגיות בכתיב חסר הניקוד".

הכתיב חסר הניקוד - דוגמאות

- הוספת וי"ו לציון קיבוץ:
חולצה.
- הוספת וי"ו לציון חולם, קמץ קטן וחטף-קמץ:
חודש, **חודשי**, **חודשים**.
- אך רק אם היה חולם במילת הבסיס:
חכמה, **תכנית**, **צהריים**.
- ובפרט לא אם כלל לא הייתה תנועת o:
פועל-פועלו, **שמר-לשמרו**.

הכתיב חסר הניקוד - דוגמאות

- מוסיפים יו"ד לציון חיריק שאין אחריו שווא נח: **סיפור**, **שנייה** , **יישום**, אבל **שמחה**, **עברית**, וכן **דיוק**.
- במילים לועזיות, היו"ד כבר בכתיב המנוקד: **היסטוריה**.
- הוספת יו"ד בנטיות תלויה בצורה היסודית: **זיכרון-זיכרונות**, **לב-לבי**.
  אבל **עיפרון-עפרונות**.
- הוספת יו"ד לציון צירי במקרים מסוימים: **תיבה**, **חירש**, אבל לרוב לא: **שרפה**, **ממד**.

הכתיב חסר הניקוד - דוגמאות

- הכפלת ו״ו עיצורית:
עוול, **הוועד**, **עכשווי**, אבל **ועד**, **קו**, **מצוות**.
- הכפלת יו״ד עיצורית:
בניין, **הייתה**, **עליי**, **קריית**-, אבל **קריה**, **מצוין**, **ילד**, **הילד**, **בנאי**, **בית**.

רשימת המילים ובנייתה

- מה יש ברשימת המילים?
- **שמות עצם.** הטיה אוטומטית ליחיד, רבים, נפרד, נסמך, כינויים.
דוגמה: **כלב**:

כלב כלב- כלבי כלבנו כלבך כלבך כלבכם כלבכן כלבו כלבה כלבן כלבם	כלבים כלבי- כלביי כלבינו כלביך כלבייך כלביכם כלביכן כלביו כלביה כלביהן כלביהם

רשימת המילים ובנייתה (המשך)

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

ירוק ירוק- ירוקים ירוקי-	ירוקה ירוקת- ירוקות ירוקות-

רשימת המילים ובנייתה (המשך)

- **פעלים.** הטיה אוטומטית לבניינים אפשריים, זמנים, שמות פעולה, וכו'.
דוגמה: **שמר**, בניין קל (רשימה מקוצרת):

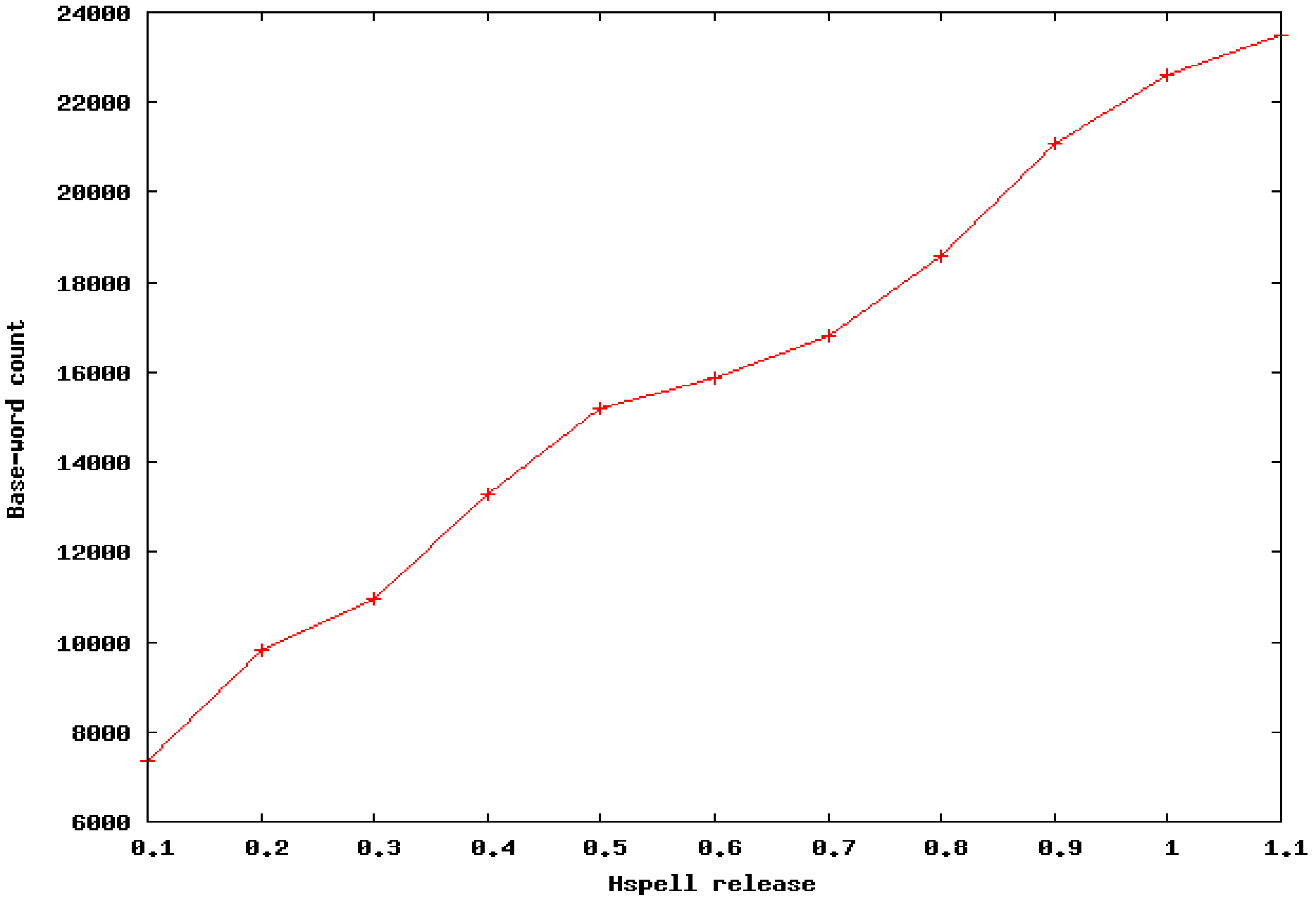
לשמור	שמרתי	שומר	אשמור	שמור	לשמרני וכו
שמירה	שמרת	שומרת	תשמור	שמרי	שמרתיו וכו
	שמרת	שומרים	תשמרי	שמרו	
	שמר	שומרות	ישמור	שמורנה	
	שמרה	שמור	תשמור		
	שמרנו	שמורה	נשמור		
	שמרתם	שמורת-	תשמרו		
	שמרתן	שמורים	תשמורנה		
	שמרו	שמורי-	ישמרו		
		שמורות			

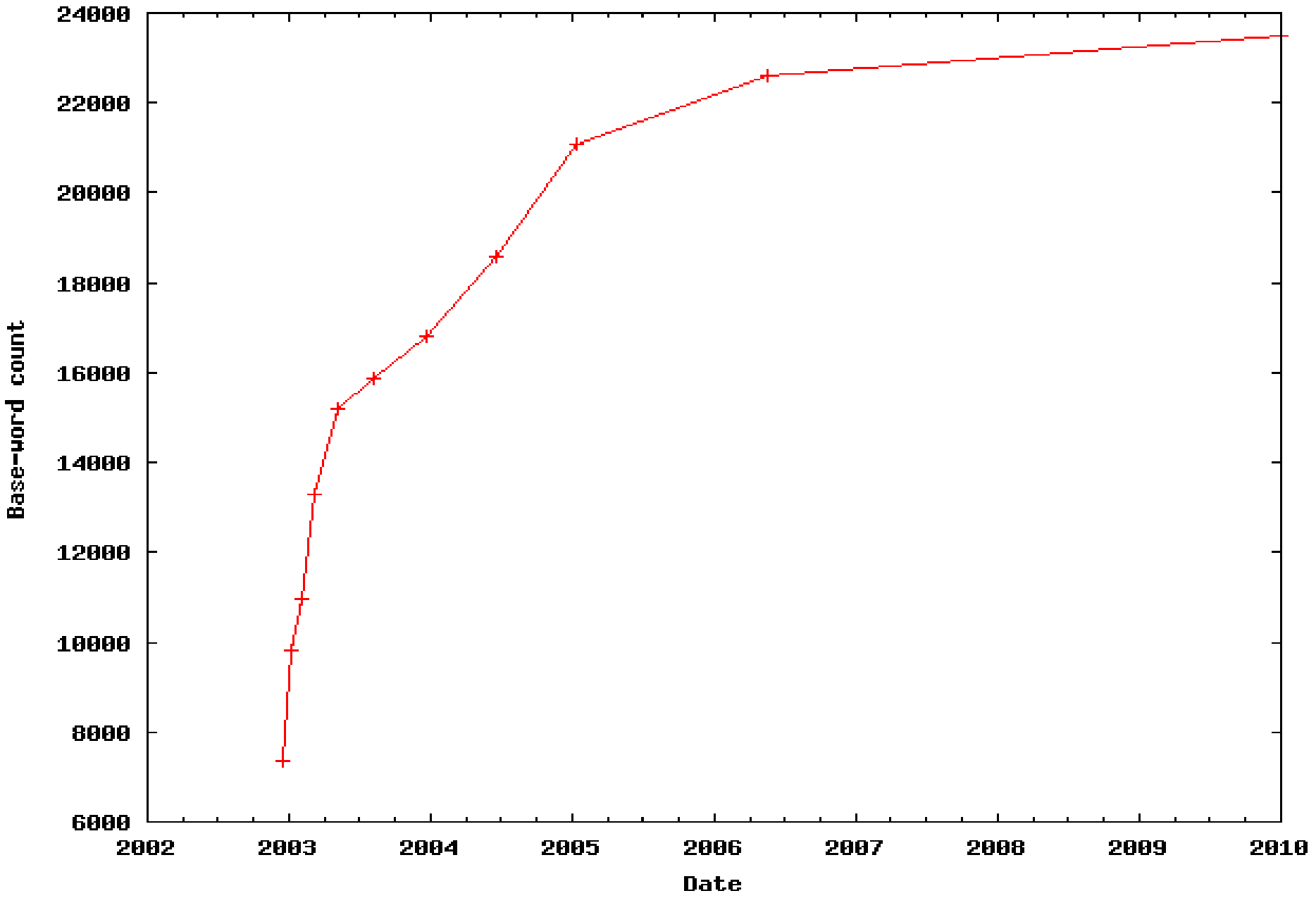
רשימת המילים ובנייתה (המשך)

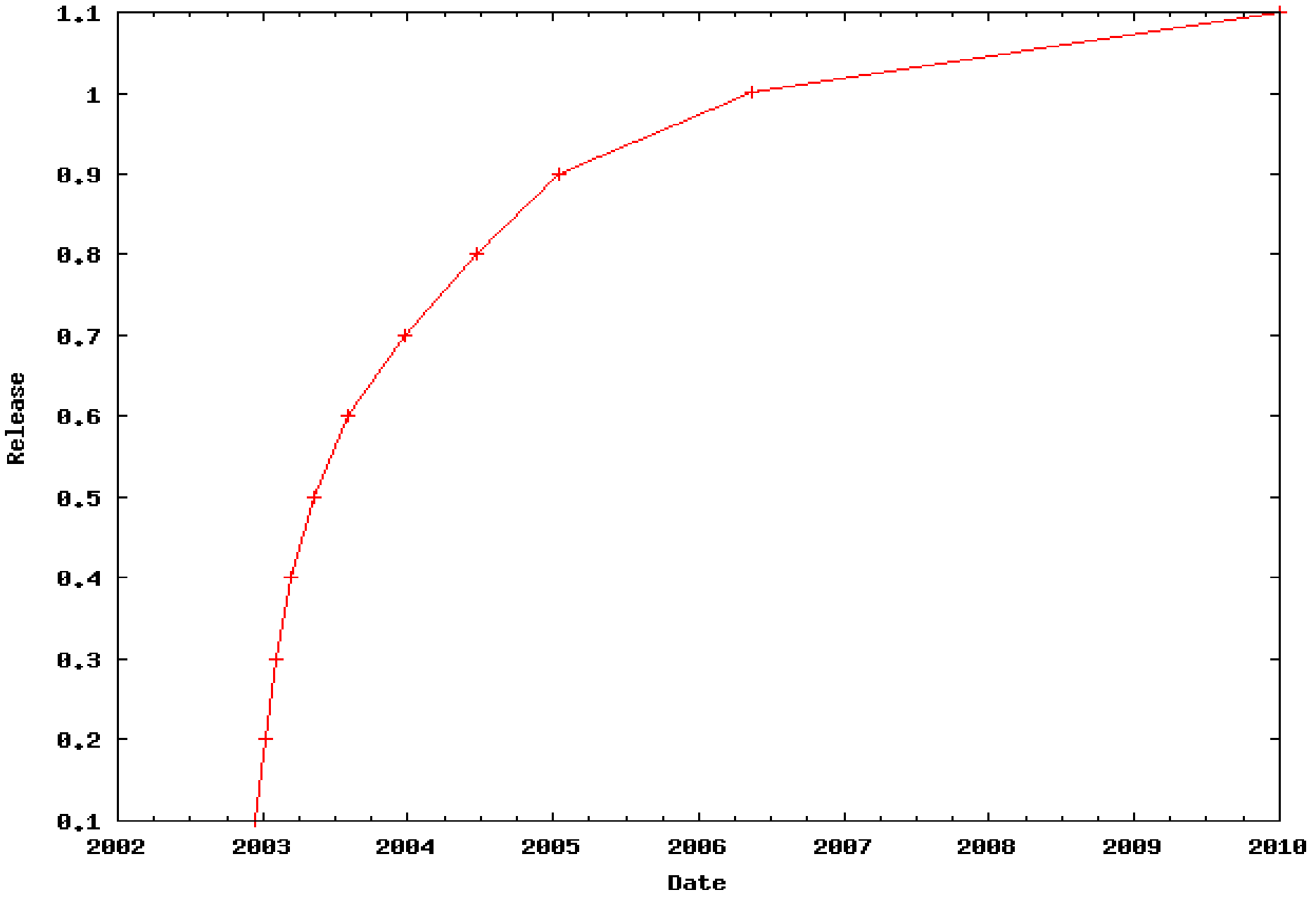
- **מילים נוספות**, ללא הטיה אוטומטית:
  - מיליות ונטיותיהן (של, שלי, ...)
  - תוארי הפועל (היכן, שלשום, הרבה, ...)
  - שמות גוף (אני, אלה, מיהו, ...)
  - שמות פרטיים של אנשים, מקומות, חודשים, אותיות וכד'.
  - מספרים (שניים, שניהם, שני)

רשימת המילים ובנייתה (המשך)

- גרסה 1.1 כללה נטיות של:
  - 12445 שמות עצם
  - 3583 שמות תואר
  - 5242 פעלים
- וכן 2216 מילים נוספות
- ובסך הכול:
  - 23,486 מילות בסיס
  - 461,326 נטיות ומילים


24000
22000
20000
18000
16000
14000
12000
10000
8000
6000
Base-word count
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
1.1
Hspell release


24000
22000
20000
18000
16000
14000
12000
10000
8000
6000
Base-word count
2002
2003
2004
2005
2006
2007
2008
2009
2010
Date


1.1
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
Release
2002
2003
2004
2005
2006
2007
2008
2009
2010
Date

הטיית שמות עצם

- התכנית wolig.pl מקבלת רשימת שמות-עצם ומטה כל אחד מהם, בעזרת רמזים שנתן המשתמש.
- ב"לוח השמות השלם" של ד"ר שאול ברקלי מופיעות כ- 350 צורות נטייה לשמות עצם מנוקדים.

הטיית שמות עצם (המשך)

- למזלנו רוב הצורות מתבררות כזהות ללא ניקוד. למשל

  כלב-כלבי דגל-דגלי

- שתי אלו מוטות באותה צורה ע"י שורת הקלט:

  כלב ע

- והפלט הוא

  כלב כלב- כלבי כלבנו כלבך כלבך כלבכם כלבכן כלבו כלבה כלבן כלבם כלבים כלבי- כלבי כלבינו כלביך כלבייך כלביכם כלביכן כלביו כלביה כלביהן כלביהם

הטיית שמות עצם (המשך)

- לעתים חייבים רמזים לנטיות הנכונות, למשל לגבי צורת הריבוי.
  - השוו: קוף-קופים עוף-עופות
  - או: שירות-שירותים חירות-חירויות
  - לשם כך חייבים לספק רמזים:
    עוף ע,ות
    שירות ע,ים
  - והפלט למשל: עוף עוף- עופי עופנו עופך עופך עופכם עופכן עופו עופה עופן עופם **עופות** עופות-עופותיי עופותינו עופותיך עופותייך עופותיכם עופותיכן עופותיו עופותיה עופותיהן עופותיהם

הטיית שמות עצם (המשך)

- צורות ריבוי נוספות:
  - משנה ע,יות
  - קצבה ע,אות
  - גרב ע,יים
  - עשן ע,יחיד
  - בת ע,רבים=בנות

הטיית שמות עצם (המשך)

- ישנם גם שמות-עצם עם כמה צורות ריבוי חוקיות:
  - חודש ע,ים,יים
  - שעה ע,ות,יים
  - קבר ע,ות,ים
  - איש ע,ים,רבים=אנשים
  - שפה ע,ות,יים,רבים=שפתות

הטיית שמות עצם (המשך)

- התכנית wolig.pl מנסה לנחש את הריבוי הנכון לפי סיומת צורת היחיד:

יחיד	רבים
מלך	מלכ**ים**
מלכ**ה**	מלכ**ות**
כותר**ת**	כותר**ות**
כמ**ות**	כמו**יות**
אחריות	-אין-

- כך כ 90% משמות העצם אינם דורשים רמז ריבוי

הטיית שמות עצם (המשך)

- כאמור, למזלנו לא מעניין אותנו אם תנועה כמו סגול, פתח או צירי משתנה בין הנטיות השונות:
  - כלב – כלבים, כלבי
  - ספר – ספרים
  - מטוס – מטוסים
- גם במקרה חיריק, כללי הכתיב חסר הניקוד לרוב לא ידרשו יו"ד:
  - ספר – ספרי
  - עז – עזים
- יוצא דופן המקרה בו ביחיד חיריק, בנטיות נעלם
  - עיפרון – עפרוני, עפרונות. עיפרון ע,ות,אבד_י

הטיית שמות עצם (המשך)

- למרבה הצער, דרושים עוד מספר רמזים. דוגמאות:
  - חנית – חניתות (לא חניות): חנית ע,שמור_ת
  - רובה – רובי (לא רובתי): רובה ע,ים,סגול_ה
  - צומת – צמתים, צומתי: צומת ע,ים,אבד_ו
  - אח, אחיו (לא אחו): אח ע,מיוחד_אח
  - שן, שיניים (לא שניים): אח ע,מיוחד_שן
  - מגורים, מגורי-, מגוריי...: מגור ע,אין_יחיד
  - גברת ע,רבים=גברות,נסמכים=גבירות

הטיית שמות עצם (המשך)

- טיפול אוטומטי בכללי הכתיב חסר הניקוד, תוך שימוש פנימי באותיות עזר:
  - קריה ע: קרעה קריה קרעת- קריית-
  - עירייה ע: עיריעה עירייה עיריעות עיריות
  - קו ע: קwקשים קו קוים
  - שוק ע,אבד_ו: שוקwשקים שוק שווקים שוקיכם שוקיכם

בודק איות חופשי

- דרגות חופש אפשריות לבודק איות:
  - קנייני ומסחרי (קוד נעול, תמורת תשלום).
  - חינמי אך קוד נעול (freeware).
  - קוד פתוח, רשימת המילים הסופית נתונה כקובץ.
    מבחינת RMS, זה חופשי. לדעתנו לא מספיק:
    איך מפתח חדש מוסיף או משנה מילים?
  - קוד פתוח, נתונה רשימת מילות בסיס וקוד נטייה פתוח.
    איך יודעים מהו תקן האיות המשותף לכל המילים?
  - קוד פתוח, מילות בסיס וקוד נטייה פתוח, ובנוסף תיעוד מפורט של תקן האיות הממומש. **זו בחירתנו.**

בודק איות חופשי

- כמובן שקוד או תוכן מועתק אינו חופשי, גם אם מדביקים עליו GPL:
- המילון פותח ללא העתקת רשימות מילים ממוצרים אחרים, מילונים, וכד'.

בודק איות חופשי

- החלטנו לשחרר את Hspell תחת ה GPL.
  האם זו הייתה החלטה נכונה?
- שירותים רשת יכולים להשתמש ב Hspell כי הם לא מפיצים תכנה. לדוגמה, gmail.
- האם ואיך מגן ה GPL עד המילון? אנחנו יודעים על לפחות מקרה אחד בו המילון שלנו (אך לא הקוד) הועתק בניגוד לתנאי ה GPL.
- תכנות שאינן GPL כמו openoffice לא יכלו להכניס את hspell. שימוש ב aspell עקף את הבעיה.
- באותו אופן, יכולה מערכת הפעלה קניינית להכיל בודק איות עברי מבוסס hspell.

מעבר לבדיקת איות – הווה ועתיד

- עד עתה דיברנו על בדיקת איות.
- מהן היכולות הלשוניות הנוספות שקיימות ב hspell כבר היום?
- מהן היכולות הלשוניות החסרות ושאולי נצטרך בעשור הבא?

מנתח צורני (מורפולוגי)

- התכנית שמייצרת את נטיות "כלב" יכולה לזכור בדיוק מדוע יצרה כל נטייה:

כלב ע,ז,יחיד
כלב ע,ז,יחיד,סמיכות
כלבי ע,ז,יחיד,של/אני
...
כלבים ע,ז,רבים
....

- כך בהינתן מילה בטקסט (נטויה, אולי עם תחיליות), נוכל לדעת את הצורה או הצורות האפשריות לנתח אותה.

מנתח צורני (מורפולוגי):

- דוגמאות מ hspell -al:

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,ז,2,יחיד,עבר) רכבת (ע,נ,יחיד)	**משטרה**: משטרה (ע,נ,יחיד) משטר (ע,ז,יחיד,של/היא) **מ+שטרה**: שטר (ע,ז,יחיד,של/היא)

מה עוד חסר?

- יכולות חדשות שלדעתי נצטרך בעשור הקרוב בעולם התכנה החופשית: (סדר חשיבות יורד)
  - תכנת הקראה (TTS) של טקסט עברי לא מנוקד.
  - בודק איות שמזהה גם מילים נכונות בהקשר לא נכון. לדוגמה, "זהו כתיב לא תיקני".
  - תכנת חיפוש טקסט עברי מדויקת, שמאפשרת למצוא את המילה שהתכוונת אליה (למשל רכבת, לא רכבתָ).
  - (עדיפות נמוכה) תכנת ניקוד אוטומטי
  - (עדיפות נמוכה מאוד) בודק איות לכתיב מנוקד.

מה עוד חסר? (המשך)

- איך לעשות זאת? לדעתי:
  - תכנה לניתוח דקדוקי בהינתן ניתוח מורפולוגי.
  - שיפור דיוק על-ידי לקסיקון עם מחלקות סמנטיות (semantic frames). לדוגמה: "שתיתי חלב".
  - הוספת ניקוד – או לפחות תנועות – ללקסיקון ולתכנת הנטייה של hspell.
  - הכרת וריאנטים נפוצים (לוא דווקא נכונים) למנתח המורפולוגי של hspell.
  - תכנת הקראה של טקסט עברי עם תנועות.
- לחלק מהיכולות לא דרושים כל השלבים.

חוסרים קטנים

- שיפור פיצול מילים ב openoffice (ראשי תיבות)
- כיווץ עץ המילים בזיכרון: מ 5 MB לכיוון 150 KB. יאפשר גם mapping במקום קריאה, ו embedded
- שיפור אלגוריתם הצעת התיקונים של hspell.
- שיפור הצעות התיקונים ב hunspell.
- עבודה על יעילות hunspell.
- ייעול המנתח הצורני (כרגע צורך 17 MB).
- שילוב המנתח הצורני ב hunspell.