שקף 1

עברית בלינוקס

צפריר כהן

tzafrir@technion.ac.il

שקף 2

תוכן ענינים

שקף 3

איך זה נכתב

מסמך זה הוכן, כמובן, על מערכת לינוקס.

שקפים אילו הוכנו ע"י תוכנת LaTeX (בעזרת התמיכה העברית לLaTeX2e)

הקבצים נערכו בעורך VIM, עם קצת עזרה מהחבילה TKTeX, אותה אפשר להוריד מ: http://psl-proxy.technion.ac.il/TKTEX

שקף 4

עברית קשה שפה

יש לנו שפה מוזרה:

- לא יכולים להשתמש באותיות רגילות
- חייבים סימונים שונים על המקלדת
- ואם זה לא מספיק - גם כותבים מימין לשמאל

למה בכלל להתאמץ לכפות אותה על המערכת שלנו?

שקף 5

למה להתאמץ

כל מי שעובד עם מחשב אישי רואה שאפשר להוסיף תמיכה עברית למערכת הפעלה.

MacOS וחלונות כבר הדגימו שזה אפשרי, אם כי במחיר של בניית כל המערכת מחדש.

בשנים האחרונות חברת מיקרוסופט הוסיפה תמיכה יותר מלאה של עברית במוצריה השונים, כחלק מיישום תקן Unicode, ללא צורך בבניה מחדש של המערכת.

אז כנראה שתמיכה עברית היא בכל זאת משהו אפשרי

שקף 6

א׳ קידודי תווים

הדרכים השונות לייצג תווים

- קידוד ASCII
- קידודי 8 ביטים
- יוניקוד

שקף 7

איך המחשב מבין עברית?

אנחנו יודעים שבשביל המחשב כל דבר הוא בסך הכל אוסף של אחדות ואפסים.

אבל למעשה כאשר עובדים עם טקסט עובדים בד"כ עם תוים (characters).

תו הוא אוסף ביטים (בד"כ 8 או 16) שמייצגים בד"כ אות אחת בטקסט.

שקף 8

בראשית היה ה-ASCII

פעם חשבו שבעזרת 7 ביטים (שנותנים 128 צירופים אפשריים) אפשר לייצג את כל התווים הנדרשים לתקשורת. לכן יצרו את סטנדרט ASCII:
American Standard Coding for Information Interchange

והוא אכן מכיל (כמעט) את כל התווים הנדרשים.

אם עובדים באנגלית בלבד.

שקף 9

קידוד 8Bit

יותר מאוחר הבינו שמכיוון שממילא עובדים במחשב עם יחידות של 8 ביטים (או כפולות שלהם) אז אפשר לקודד תו ע"י 8 ביטים במקום שבעה. בדרך זו מתפנים עוד 128 קידודים אפשריים (התווים בעלי המספרים 128-255), ואפשר לנצל אותם כדי לשים שם כל מיני תווים שימושיים: קווים, אותיות יווניות, אותיות לטיניות עם סימני מבטא (כגון ã), וכו'.

אפשר גם לשים שם תווים משפה אחרת, כגון תאילנדית או עברית.

הבעיה היא שכל אחד רצה לשים שם תווים אחרים ובמקומות אחרים.

שקף 10

קידודי 8bit (המשך)

מכיוון שכך - כדי לדעת איך הטקסט נראה צריך לדעת לא רק מהו הטקסט עצמו, אלא גם באיזה דף קוד (codepage) הטקסט כתוב.

וכאשר שולחים קובץ למדפסת, או טקסט בדואר אלקטרוני, צריך גם שם להגדיר מראש באיזה קידוד בדיוק משתמשים. אחרת אם אני אשלח את הטקסט הבא:

And Some Hebrew קצת אנגלית

אני עלול לקבל:

And Some Hebrew ãêíäúé ‘aóü

קידודי 8bit (המשך)

נוצרו די הרבה קידודים שונים. חלקם גם כללו עברית במקומות שונים. לבסוף רובם מיקמו את תווי העברית בתחום 224-250 (היוצא היחיד מן הכלל הזה הוא cp862 שנוצר ע"י IBM לצורך התמיכה העברית בדוס).

ISO8859 הוא סטנדרט של ISO שמגדיר כמה קידודים סטנדרטיים לכמה שפות. מביניהם **ISO8859-8**. לפעמים קוראים לו פשוט **iso8**.

windows1255 הוא זה שבו משתמשות תוכנות חלונות השונות. די דומה ל-iso8859-8, אולם מוסיף עוד תווים, ביניהם סימני ניקוד (תווי ניקוד הם תווים בעלי רוחב 0, ולכן מוצגים באותו מקום כמו התו הקודם).

למעשה - אם אתם משתמשים רק בתווי ASCII ובתווים עבריים - הקידודים הללו זהים.

שקף 11

אבל מה אם אני רוצה גם עברית וגם רוסית?

אכן יש בעיה. כל הקידודים הללו טובים למקרה שבו רוצים לעבוד עם אנגלית ועם עוד שפה אחת. אולם הם אינם מועילים במיוחד אם רוצים לעבוד עם עוד שפה.

מה עושים אז?

בשביל לפתור את הבעיה הזו הוצע תקן Unicode, שנועד לתת קידוד אחיד לכל השפות. ב-Unicode מציע קידוד אחיד שכולל את כל השפות. ואפילו קצת יותר מזה ...

שקף 12

שקף 13

Unicode

תקן יוניקוד, הידוע גם כ-ISO-10646-1, מנסה לדאוג שיהיה מקום להכל: כל סימן פיסוק אפשרי בכל ווריציה אפשרית, וכל שפה: לטינית, רוסית, ארמנית, קירגיזית, אורדו. אפילו עברית.

בתקן יוניקוד סימני העברית נמצאים באזור U+0590-U+05FF (בתוי יוניקוד מקובל להשתמש בספרות הקסדצימליות לסימון קידודי התווים). האותיות עצמן נמצאות בטווח U+05D0-U+05EA (רוב התווים האחרים הם סימני ניקוד וטעמים).

שקף 14

ב׳ גופנים עבריים ב-X

- איך קוראים לגופנים ב-מערכת X
- הגדרת גופן עברי ביישומים (ידנית)

שקף 15

שמות הגופנים ב-X

כאשר שם הגופן מתחיל ב-'-' זה סימן שיש לו שם די ארוך, שמפרט הרבה מתכונותיו. לדוגמה:

-misc-fixed-medium-r-semicondensed–13-100-100-100-c-60-iso8859-8

לפעמים מספיק לפרט רק חלק מהתכונות של הגופן, ואז שמים כוכביות בשאר השדות:

```
-misc-fixed-*-*-*-*-*-*-*-*-*-*-iso8859-8
```

זוהי בקשה מהשרת למצוא פונט מסויים שיתאים לתאור (הכוכבית היא כאן wildcard)

[הדגמת xfontsel עם -*-arial-*-iso8859-8]

xfontsel אינה תוכנה נוחה במיוחד לעבודה. גם גנום וגם KDE באים עם סוקרי גופנים משלהם, שגם בד"כ משמשים לבחירת גופנים.

שקף 16

שינוי גופנים ביישומים

לא כל העולם משתמש בעברית. ברוב העולם משתמשים בכל מיני תווים אחרים איפה שאנחנו שמים את העברית. לכן כאשר אנו רוצים לראות עברית, אנו רואים ג'יבריש.

[הדגמת מסמך טקסט עברי-ג'יבריש ב-gedit.]

אם כבר יש לנו במערכת גופנים (פונטים) עבריים, הדרך הפשוטה ביותר היא להגיד לאותו יישום להשתמש בהם. בד"כ נפגוש גופנים ב-קידודים שונים. כאמור, הגופנים העבריים יופיעו בד"כ בקידוד *ISO-8859-8*

[הדגמת שינוי הגופן ב-gedit.]

נשים לב שהעובדה שהמסמך מוצג כעת בקידוד עברי לא גורמת לכך שהעברית תוצג בצורה הפוכה. נתייחס לכך בהרחבה בהמשך (עברית דו-כיוונית).

כמוכן, היה רצוי שנגדיר שלא נצטרך להגדיר לכל ישום בנפרד לעבוד עם גופנים עבריים. אנו נראה בהמשך כמה דרכי הגדרה יותר כלליות.

xterm בעברית: מה שעובד תמיד

כמעט בכל סביבת X מוגדר לפחות גופן אחד עברי אחד. בפרט עם כל הפצת X11 סטנדרטים מגיע הגופן misc-fixed בקידוד iso8859-8.

להרבה מהישומים בסביבת X אפשר לקבוע את הגופן שבו ישתמש היישום בעזרת האופציה -fn *font name*. לדוגמה:

שקף 17

```
xterm -fn '-misc-fixed-medium-r-normal-*-*-*-*-*-*-*-iso8859-8'
```

השתמשתי ב-'*' עבור כל פרמרטרי הגודל, כי לא היה אכפת לי איזה גודל ייבחר. הייתי יכול לכתוב במפורש:

```
xterm -fn
-misc-fixed-medium-r-normal--13-100-100-100-c-80-iso8859-8
```

זה באמת די ארוך לכתוב את כל השם הזה. לכן תמצאו כמעט בכל מערכת לפחות שני כינויים aliases: heb6x13, heb8x13. ולכן, אם אתם רוצים לפתוח חלון xterm עם גופן עברי:

```
xterm -fn heb8x13
```

ג׳ הוספת גופנים

שקף 18

שרת ה-X אחראי לספק לנו גופנים. אם אתרע מזלינו והוא לא מספק לנו גופנים עבריים, אנו נאלץ להוסיפם בעצמנו.

שרת ה-X יכול לקבל את הגופנים משרת חיצוני. זוהי גם הקונפיגורציה הסטנדרטית בהפצות רד-האט ומנדרייק.

- תיקיות גופנים
- FontPath
- שרתי גופנים
- שרת של רד-האט ומנדרייק

תיקיות גופנים

ישנן דרכים רבות לשמור גופנים: True Type, Type1, Type3, וכו'. בכולם הגופנים נשמרים בקבצים.

שרת ה-X שומר אצלו רשימה של תיקיות שבהן יש גופנים.

שקף 19

אם נסתכל על שמות הקבצים בתיקיה כזו, נשים לב לשני קבצים בעלי שמות לא מופרעים לחלוטין: fonts.dir ו-fonts.alias (לעיתים יש גם fonts.scale).

ב-fonts.dir רשום לגבי כל קובץ, מהם שמות הגופנים שאותם הוא מספק.

ב-fonts.alias אפשר להוסיף כינויים: אפשר להוסיף שם נוסף לגופן קיים.

(אל פחד: fonts.dir נוצר אוטומטית. נראה איך בהמשך. fonts.alias לא הכרחי, אולם הוא יכול להיות די מועיל).

הוספת תיקיות גופנים

שקף 20

אם לא קיים קובץ fonts.dir צריך לייצרו. הפקודה `mkfontdir` יוצרת את הקובץ הזה מהקבצים הקיימים (לפעמים היא משתמשת גם בקובץ fonts.scale. לחילופין, אפשר ליצור קובץ כזה ידנית (או להביא אחד מוכן).

כעת נגיד לשרת ה-X לטעון אוטומטית את הגופנים מתיקיה זו. אפשר להריץ כל פעם xset בזמן אתחול X, אבל זה מסורבל. דרך יותר נורמלית היא להוסיף את התיקיה לרשימת תיקיות הגופנים שבקובץ ההגדרות של שרת ה-X: **XF86Config**.

בקובץ זה ישנה שורה **FontPath**. לפני העריכה:

```
FontPath "/usr/lib/X11/fonts/misc;unix/:-1"
```

לאחר העריכה:

FontPath "/usr/lib/X11/fonts/misc;unix/:-1;/usr/share/fonts/newdir"

נשים לב שהגדרות אילו יכנסו לתוקף רק עם הפעלתו מחדש של שרת ה-X

שקף 21

שרתי גופנים עצמאיים

שרת X יודע להשתמש גם בשרתי גופנים (font servers) חיצוניים. שרתי גופנים אילו יכולים לרוץ על אותו מחשב או אף על מחשב שונה ברשת. שרת גופנים הוא עוד מקור אפשרי לגופנים, כמו תיקיית גופנים, ולכן יכול להופיע ב-FontPath של שרת X.

עבור שרת שהתקשורת איתו היא דרך tcp, והוא רץ על המחשב fs.foo.co.il בפורט 7100 נרשום: tcp/fs.foo.co.il:7100

במצב שיש שרת גופנים מקומי שלא מדבר עם הרשת, ומדבר עם השרת המקומי רק דרך מה שנקרא unix domain socket, אנו רושמים אותו ב-FontPath כ-unix/:-1

שקף 22

שרתי גופנים עצמאיים (המשך)

כאשר עובדים עם שרת פונטים חיצוני הוא יכול לשנות את רשימת הגופנים שהוא מציע בלי להודיע לשרת ה-X. אם שרת הפונטים פתאום יבקש משרת הגופנים גופן שזה האחרון כבר לא מספק, יכולות להגרם בעיות (במקרה הגרוע אף תקיעה של שרת ה-X).

כדי להמנע מהבעיות הללו, אם שיניתם את רשימת הגופנים ששרת גופנים מספק (בעיקר אם הורדתם גופן) ושרת ה-X רץ, מומלץ להריץ שם:

```
xset fp rehash
```

ליתר בטחון, כדי להגיד לשרת ה-X לעדכן את רשימת הגופנים שלו.

שקף 23

שרת הגופנים של רד-האט

השימוש בשרת גופנים נפרד הוא די נפוץ בהפצות הלינוקס. כמו אפשר היה לראות שאצלי ה-FontPath בקובץ XF86Config כלל רק את שרת הגופנים.

נראה כאן איך מוסיפים גופני True-Type לשרת הגופנים של רד-האט. זוהי גרסה של רד-האט לשרת הגופנים הסטנדרטי של X. גם מנדרייק משתמשת באותו שרת.

שקף 24

שרת הגופנים של רד-האט (המשך)

יצירת תיקיית גופנים חדשה

1. יצירת התיקיה:

```
mkdir fonts_dir
```

2. יצירת קובץ fonts.dir:

```
cd fonts
ttmkfdir > fonts_dir
```

3. עריכת /etc/X11/fs/config---הוספת *fonts_dir* ל-catalog
4. טעינת הקונפיגורציה של שרת הגופנים מחדש:

```
/etc/rc.d/init.d/xfs restart
```

5. (אם שרת ה-X רץ) לעדכן את רשימת הגופנים:

```
xset fp rehash
```

שרת הגופנים של רד-האט (המשך)

עדכון תיקיית גופנים קיימת

לאחר כל הוספת או הורדת גופן True-Type מהתיקיה צריך לעדכן את המערכת. כאן מספיקים רק השלבים הבאים:

שקף 25

1. יצירת קובץ fonts.dir:

```
cd fonts
ttmkfdir > fonts_dir
```

2. טעינת הקונפיגורציה של שרת הגופנים מחדש:

```
/etc/rc.d/init.d/xfs restart
```

3. (אם שרת ה-X רץ) לעדכן את רשימת הגופנים:

```
xset fp rehash
```

שקף 26

ד׳ עברית ״הפוכה״

- ״עברית ויזואלית״ ו-״עברית לוגית״
- gtkbidi

שקף 27

"עברית ויזואלית" ו-"עברית לוגית"

- עברית, כמו ערבית, נכתבת הפוך משאר השפות
- לפני הצגה או הדפסה של טקסט עברי צריך "להפוך" אותו. אבל זה לא כל-כך פשוט
- ההיפוך הזה לא יכול להתבצע מוקדם מדי. לדוגמה: שבירת שורות
- הטקסט לאחר ההיפוך נקרא "עברית ויזואלית". לא נוח לעבוד איתו. עובדים בד"כ עם טקסט רגיל ("עברית לוגית")

התכונה הזו משותפת לעברית ולשפות דו-כיווניות (BIDIrectional) אחרות, כגון ערבית. תקן יוניקוד מגדיר בדיוק איך לבצע את ההמרה מעברית לוגית לעברית ויזואלית.

שקף 28

gtkbidi

- משתמשת בספריית fribidi
- מחליפה פונקציות של ספריית gtk המממשות תיבת טקסט בתיבת טקסט עם תמיכה בעברית דו-כיוונית
- עובדת בצורה סבירה ומעלה עם רוב ישומי gtk
- אבל לא עובדת עם gtk 1.2.8 ומעלה

השימוש:

gtkbidi *program_name* [*parameters...*]

מידע נוסף:

http://linux.org.il/faq/?file=130

שקף 29

ה׳ מיפוי מקלדת ב-X

איך גורמים לכך שלחיצה על 'A' תשלח 'ש'?

- Xkb
- הגדרת מקלדת עברית עם Xkb
- אלטרנטיבות

שקף 30

X KeyBoard extention - Xkb

- הרחבה לפרוטוקול X כדי לטפל יותר טוב במיפוי מקלדת
- מאפשרת עבודה עם עד ארבע פרישות מקלדת (groups)
- בעלת קונפיגורציה גמישה וחזקה. אבל...
- סובלת מחוסר תיעוד

אופן השימוש: מבקשים מהשרת לקמפל מפת מקשים מתוך תאור סטנדרטי. ישנן שלוש רמות תאור:

1. מקלדות מוכנות: Keymaps
2. תיאור לפי חוקים (Rules): Model (לדוגמה pc101, pc104), Layout (לדוגמה: ישראלי: il, אנגלית-ארה"ב: en_US) ו-Options (אופציות נוספות)
3. תיאור בעזרת מיפוי: עבודה ברמה יותר נמוכה: symbols, compat, keycodes, types

Xkb - הגדרת מקלדת עברית

ראשית כל צריך את קובץ ה-symbols עבור מקלדת ישראלית:

ftp://linux.org.il/pub/Hebrew/Keymaps/il.symbols.xkb

את הקובץ הזה צריך להעתיק ל: /usr/lib/X11/xkb/symbols/il

כעת אפשר להוסיף את השורות הבאות לקובץ XF86Config, בחלק Keyboard

שקף 31

לזכור לבטל את כל השורות של XkbModel, XkbLayout ו-XkbOptions.

ולהוסיף:

```
XkbCompat  "group_led"
XkbSymbols "us(pc104)+il+group(switch)+group(shift_toggle)"
```

שימו לב שב-XFree4 התחביר קצת השתנה: החלק Keyboard הפך להיות InputDevice, והתחביר קצת השתנה. לדוגמה: במקום `'XkbCompat "group_led"'` יופיע `'Options "Compat" "group_led"'`

כלים אחרים למיפוי המקלדת

כמה הערות קצרות על כלים אחרים המשמשים למיפוי מקלדת:

- xkbcomp - מאוד לא נוח
- setxmodmap - כלי הרבה יותר נוח לשינוי המיפוי משורת הפקודה
- Xmodmap - פרימיטיבי ולא נוח, אבל עובד תמיד
- kikbd - שבא עם kde1 - צריך רק לעדכן מיפוי מקשים
- GKB (של גנום) - הגרסה הנוכחית משתמשת למעשה ב-setxkbmap
- WMKeyboard - צריך לעדכן מיפוי

שקף 32

שקף 33

ו׳ לוקליזציה

לוקליזציה של המערכת: שינוי התנהגות המערכת בהתאם לשפה.

- למה זה טוב?
- איך משנים?
- איך זה משפיע על מקלדת ב-X?

שקף 34

Locales: לוקליזציה של תוכניות

יש הרבה דברים במערכת שהיינו רוצים שיתנהגו בצורה שונה לפי השפה של המשתמש.

[הדגמת date בשפות שונות]

מה זה? אם אני מצהיר שהשפה (LANG) שלי היא עברית (he) המערכת נותנת תשובה שונה לשאלה "מה השעה" אם אני מצהיר שהשפה שלי היא ברירת המחדל (C).

כמוכן השפה משפיעה על סיווג תווים: ישנם תווים שנחשבים "חוקיים" בחלק מהשפות, אך לא באחרות:

[הדגמת less עם קובץ בעברית בשפות שונות]

סביבת X משתמשת בלוקליזציה בצורה חזקה. לדוגמה: לכל ישום X אפשר להוסיף קובץ הגדרות שתלוי בשפה, ולכן לשנות את התנהגותו כתלות בשפה.

שקף 35

איך משנים שפה

ראינו שהשפה נקבעת עפ"י המשתנה LANG. כדי לשנות אותו צריך לערוך את אחד מקובצי האתחול.

כמוכן אפשר לשנות רק חלק מהדברים, בעזרת המשתנים LC_CTYPE, LC_COLLATE, LC_MESSAGES, LC_MONETARY, LC_NUMERIC, LC_TIME.

לדוגמה: אם אני רוצה רק שסוג התווים יהיה בעברית, אני אגדיר:

```
LC_CTYPE=he
```

ואם אני רוצה שהכל, למעט ממשק המשתמש, יהיה בעברית

```
LANG=he
LC_MESSAGES=C
```

לפרטים מלאים: locale (7)

שקף 36

הבעיה: יותר משפת קלט אחת ב-X

קלט לתוכניות X מיוצג כ-"תווי X". בהם הוקצה מקום גם לתווים העבריים

כשישום מקבל תווים כאלו הוא בד"כ משתמש בפונקציה מיוחדת של X להפיכתם לתווים רגילים.

פונקציה זו תוכננה גרוע: היא מתעלמת מתווים שאינם חוקיים לפי השפה הנוכחית.

לכן אם השפה שלי תהיה עברית, אני אוכל להקליד רק תווי ASCII ותווים עבריים (ולא אוכל להקליד תווים רוסיים, לדוגמה, כי התוכניות יתעלמו מהם).

שקף 37

הבעיה: יותר משפת קלט אחת ב-X (המשך)

איך הסתדרו עד היום?

מיפו את התווים העבריים לתחום שלהם ב-ISO-8859-8. למרבה הפלא זה עובד.

כאשר השפה אינה עברית.

חוץ מזה, יותר ויותר ישומים (לדוגמה: מוזילה) מתחילים לעשות בעצמם את ההמרה, ולכן לא יתייחסו למיפוי "8bit" כתווים עבריים.

לכן לא נעבוד איתו.

שקף 38

ז׳ עברית בנטסקייפ

- "הפאטצ׳ הטורקי"
- שינוי פונטים: מה שאי אפשר לשנות דרך edit→preferences
- ואפילו עברית לוגית

שקף 39

הפאטצ' הטורקי

מתכנתי נטסקייפ החליטו משום מה לנטרל את התמיכה בדף הקוד העברי בנטסקייפ. אבל מסתבר שאפשר לערוך את הקובץ הבינארי של נטסקייפ ולהוסיף חזרה את התמיכה הזו, על חשבון התמיכה בטורקית.

הדבר יאפשר לבחור במפורש את הגופן לעברית (ללא צורך ב-"web fonts" ושאר התחכמויות.

הביצוע פשוט. מורידים את הסקריפט מ:

ftp://linux.org.il/pub/Hebrew/Netscape/ns_turkish_patch.sh

ומריצים (כ-root).

שקף 40

גופנים עבריים

גם לאחר שהפעלנו את ההפאטצ', נשארו בעיות:

טפסים: משתמשים משום מה בגופן של latin1

תפריטים (בפרט שמות סימניות): ניתן לשנות דרך X resources

biditext - לראות (כמעט) עברית לוגית

biditext עובד על עיקרון די דומה ל-gtkbidi, אבל ברמה יותר נמוכה ועל קוד הרבה יותר יציב (ולכן לא נשבר כל חצי שנה).

הוא עובד כמעט עם כל ישומי X, אבל קצת יותר מוגבל. לדוגמה: הוא לא יעבוד טוב עם עורך טקסט.

שקף 41

כדי לנטרל אותו, מוחקים את הקובץ $HOME/.rev (או BIDITEXT_FILENAME. יש גם ממשק גרפי פשוט לזה (r2l)

השימוש:

```
biditext program_niame [parameters...]
```

פרטים נוספים וקבצים:

http://www.iglu.org.il/faq/?file=129

ח' LyX - התמלילן העברי הראשון

LyX הוא ממשק גרפי למערכת LaTeX. להבדיל מ-LaTeX, מאוד קל ללמוד לעבוד איתו. בדומה ל-LaTeX- הוא מאוד מועיל.

שקף 42

התמיכה העברית נוספה בגירסה 1.1.5. שימו לב ש-kLyX אינו כולל תמיכה בעברית.

כדי שהתמיכה העברית תעבוד נדרשת, בנוסף לחבילת tetex, גם חבילה של תמיכה עברית ב-LaTeX 2_ε

מידע נוסף:

http://www.iglu.org.il/faq/?file=125