(לָא)טֵךְ רַב־לְשׁוֹנִי

Multilingual (LA)TEX (لَا)تِخْ بِٱللُّغَاتِ ٱلْمُتَعَدِّدَةِ

Ron Artstein רון ארטשטיין رون ارتشتين

Technion טכניון تخنيون

artstein@cs.technion.ac.il

8 במרץ 2004 ט״ו באדר, תשס״ד

מה זה TEX

תוכנה לסידור דפוס.

- פותחה ע"י Donald Knuth ותלמידיו בסטנפורד, כתגובה להידרדרות איכות הדפוס עם המעבר לסידור דפוס ממוחשב.
- פיתוח החל ב־1977, שחרור ב־1978, יציבה מ־1982, הוקפאה ב־1990 (מאז: תיקוני באגים בלבד).
- **יציבה מאוד:** מספר הגירסה מתכנס ל־π, כל שינוי מוסיף ספרה עשרונית. גירסה 3.14159 ב־1995, 3.141592 ב־2002. מי שמוצא באג מקבל המחאה בסך 327.68$ מקנות'.
- ב־public domain. מתועדת במלואה. כתובה ב־Web (Pascal עם תיעוד), מימוש ב־C, מימושים נוספים.
- קלט: טקסט. פלט: DVI.
- שפת תיכנות. ניתנת לקונפיגורציה ולהרחבה. כמעט לא עובדים מול TEX ישירות. הרחבות רבות, ברמות שונות של נגישות, אמינות ותיעוד.

הכוח של TEX

יש שני דברים ש־TEX יודעת לעשות טוב מאד.

- סידור דפוס מתימטי לפי קלט מעין־לוגי.

$$\int \frac{\sin^3 x}{x}\,dx \qquad \$\$\int{\sin^3x\over x}\,dx\$\$$$

שימו לב למיקום המדויק של האינדקסים: $\frac{x^2}{y^2}$ Na_2 Na_2^+

מיקום הסימנים בפלט נקבע לפי נוסחה המתחשבת ב־22 פרמטרים.

- יישור ומיקוף באופטימיזציה גלובאלית ברמת הפיסקה.

כתבתי הרצאה על סדר דפוס רב־לשוני.

כתבתי הרצאה על סדר דפוס רב־לשוני.

מקום חיתוך השורות (והמילים, אם יש צורך) נקבע ע"י שקלול מספר רב של פרמטרים המחושבים על הפיסקה כולה.

ההקפאה של TEX

תוכנת TEX מוקפאת, ולא מתבצע בה כל פיתוח. התוכנה עצמה חופשית, אבל Knuth שמר לעצמו את הזכויות על השם TEX. מותר ליישם מחדש את התוכנה; אחד התנאים הנדרשים על מנת לקרוא ליישום החדש בשם TEX הוא שהיישום ייתן פלט זהה לתוכנה המקורית בריצה על הקובץ `trip.tex`, המנצל את כל שורות הקוד בתוכנה המקורית.

פיתוח מתאפשר בשני אופנים:

- כתיבת front ends (חבילות מאקרו) ו־back ends (דרייברים) חדשים.
- תוכנה חדשה על בסיס TEX, בתנאי שניתן לה שם אחר. דוגמאות:

  TEX--XET ,TEX-XET (גירסאות דו־כיווניות),

  $\mathcal{NTS}$,ϵ-TEX (הרחבות),

  pdfTEX ,pdfϵ-TEX (פלט PDF),

  Ω ,TAUX (מערכות מבוססות־יוניקוד).

פורמטים

TEX מכירה כ־300 פקודות בסיסיות, המאפשרות סידור דפוס ברמה של הדף המודפס; כמו־כן קיימת אפשרות להגדיר פקודות חדשות על בסיס הפקודות הקיימות, כדי להשיג רמה גבוהה יותר של אוטומציה.

ניתן להגדיר פקודות חדשות בזמן ריצה, או לשמור אוסף של הגדרות ב**פורמט** לטעינה מהירה בזמן האתחול של TEX. בפועל, כמעט תמיד משתמשים בפורמט כלשהו.

Knuth הגדיר פורמט בשם `plain` הכולל כ־600 פקודות, ומיועד לתדפיסים מקדימים של מאמרים במתימטיקה. הוא גם הגדיר פורמטים נוספים עבור הספרים שכתב ופרוייקטים אחרים.

פורמטים נפוצים אחרים:

- $\mathcal{AMS}$-TEX לפרסומים של האגודה האמריקאית למתימטיקה
- ConTEXt התומך ב־XML
- LATEX

מה זה LaTeX

פורמט של TeX, ככל הנראה הפורמט הנפוץ ביותר.

- מעודד שימוש בסימון לוגי במקום סימון ויזואלי.
- מנגנונים מפותחים לסידור אוטומטי של מראי מקומות, הערות שוליים, תוכן עניינים, רשימת מקורות, אינדקס, טבלאות, איורים, ורשימות.
- גירסת LaTeX 2ε (1994 ואילך) קובעת מנגנון אחיד להרחבות באמצעות "חבילות" שנטענות לפי הצורך. יש חבילות שתומכות בגופנים, בצבעים ובגרפיקה, וחבילות רבות נוספות ליישומים מיוחדים.
- תמיכה רב־לשונית נעשית גם היא באמצעות חבילות.

פיתוח ותיכנות יכולים להתבצע ברמת ה־TeX או ברמת ה־LaTeX.

קידודי קלט

ל־TEX יש מנגנון שממפה את ייצוג התווים של קבצי הקלט לייצוג פנימי אחיד. הדבר מאפשר מימוש אחיד על מערכות בעלות דפי קוד שונים.

שימוש במנגנון המיפוי דורש פורמט חדש עבור כל דף קוד, ולכן הוא איננו מתאים למסמכים המשלבים דפי קוד שונים, או למערכת כמו LATEX אשר קוראת את כל ההרחבות באמצעות פורמט אחד.

השימוש בדפי קוד מרובים מתאפשר ב־LATEX באמצעות החבילה `inputenc`.

```
\usepackage[code]{inputenc}
\inputencoding{code}
```

החבילה הופכת את כל תווי הקלט בטווח 0x80–0xFF לתווים **פעילים**, דהיינו כל תו מהווה פקודה. הקובץ `code.def` מגדיר את המשמעות של כל תו קלט כזה.

החבילה אינה משפיעה על תווי קלט בתחום 0x20–0x7F.

החבילה מונעת הגדרת פקודות ששמותיהן כוללים תווים בטווח הגבוה.

החבילה babel

מיועדת בעיקר עבור שפות הנכתבות בכתב הלטיני ובכתבים עם מאפיינים דומים: מספר מצומצם של אותיות, ניתוח קונטקסטואלי מינימלי, האותיות נכתבות זו אחר זו ברצף.

```
\usepackage[lanaguage1,language2,...]{babel}
```

מנגנון המעבר משפה לשפה:

```
\selectlanguage{language}
\begin{otherlanguage}{language} ... \end{otherlanguage}
\foreignlanguage{language}{...}
```

בחירה בשפה מסוימת קוראת את הקובץ *language*.ldf המכיל את ההגדרות הרלבנטיות לשפה. הגדרות אלה הן בעיקר:

- טקסט אוטומטי
- קיצורי קלט ומוסכמות דפוס ייחודיות
- מיקוף

טקסט אוטומטי

פקודות רבות ב־LaTeX יוצרות טקסט שמשתנה משפה לשפה. למשל, הפקודה `\tableofcontents` צריכה ליצור טקסטים שונים בהולנדית (Inhoudsopgave), בפורטוגזית (Conteúdo), בהונגרית (Tartalomjegyzék), ובוולשית (Cynnwys). קבצי השפה השונים מכילים את הטקסטים המתאימים.

קיצורי קלט

שפות מסויימות דורשות פקודות די ארוכות להדפסה שגרתית. למשל בהונגרית, הכפלת עיצור הנכתב כצירוף של שני תווים מסומנת על־ידי הכפלת התו הראשון (**gallyak**), ובחלוקה לשתי שורות על־ידי הכפלת הצירוף כולו (**galy-lyak**).

ב־LaTeX יש לכתוב את המילה כך: `ga\discretionary{ly-}{}{l}lyak`

התמיכה ההונגרית מאפשרת את הקלט הבא: `ga'llyak`

מוסכמות דפוס

בשפות שונות נתקבעו מוסכמות שונות לגבי הטקסט המודפס. באנגלית מקובל להצמיד את כל סימני הפיסוק לטקסט, ובצרפתית מקובל להשאיר רווח קטן לפני סימני הפיסוק הכפולים ; : ? !

She said: What's this?
Elle a dit : Qu'est-ce que c'est ?

קבצי הגדרת השפה מגדירים את ההתנהגות הרצויה.

מיקוף

המיקוף הנכון נקבע בין היתר על־פי השפה.

אנגלית	sig-nal	[ˈsɪg.nl̩]
צרפתית	si-gnal	[siˈɲal]

בחירת השפה טוענת קובץ מיקוף מתאים.

כתבים לא לטיניים

babel תומכת במספר שפות הנכתבות בכתבים לא לטיניים. בנוסף להתאמות שהזכרנו, כתבים אלה דורשים קידודי קלט וגופנים מתאימים.

התמיכה ביוונית מבוססת על תעתיק לטיני: הקלט נכתב בתווים לטיניים, ומתורגם לפלט ביוונית. צורת הסיגמא נקבעת לפי ניתוח קונטקסטואלי.

Παναθιναϊκός

```
\foreignlanguage{greek}{Panajhna"ik'os}
```

התמיכה בשפות הנכתבות בכתב הקירילי מתבצעת ע"י קריאת הקלט בקידוד מקורי (למשל iso-8895-5 או koi8-r) באמצעות החבילה inputenc. תווי הקלט בקידודים השונים מתורגמים לפקודות אחידות, כדי לאפשר תמיכה שאיננה תלויה בקידוד. (קיימת גם תמיכה בתעתיק לטיני)

Динамо Москва

```
\foreignlanguage{russian}
{\CYRD\cyri\cyrn\cyra\cyrm\cyro\␣
 \CYRM\cyro\cyrs\cyrk\cyrv\cyra}
```

ויש גם תמיכה בעברית...

כתבים הודיים

כתב דבאנאגארי (देवनागरी) משמש בשפות סנסקריט (שפת כתבי הקודש העתיקים), הינדי (366 מיליון דוברים), מראתי (68 מיליון), נפאלי (16 מיליון) ושפות נוספות. בהודו נפוצים עוד לפחות 11 כתבים בעלי מאפיינים דומים.

סימן עיצור נהגה עם תנועה משתמעת *a*: क *ka*

סימן ה־virāma מבטל את התנועה המשתמעת: क *ka* + ् = क् *k*

סימן תנועה מבטל את התנועה המשתמעת; מיקומו ביחס לעיצור:

מעל: क *ka* + े *e* = के *ke* משמאל: क *ka* + ि *i* = कि *ki*

מתחת: क *ka* + ु *u* = कु *ku* מימין: क *ka* + ो *o* = को *ko*

רצף של שני עיצורים או יותר נכתב בתו מקושר, בצורה מסוגננת או מפושטת.

क् *k* + क *ka* = क्क / ॼ *kka* क् *k* + त *ta* = क्त / क्त *kta*

त् *t* + न *na* = त्न / त्न *tna* प् *p* + ल *la* = प्ल / प्ल *pla*

ष् *ṣ* + ट *ṭa* = ष्ट / ष्ट *ṣṭa* क् *k* + त् *t* + र् *r* + य *ya* = क्त्र्य *ktrya*

devnag ו־dev (פיתוח: Frans Velthius)

גופנים ב־TEX יכולים להכיל לכל היותר 256 תווים; זה לא מספיק בשביל כל צורות הקישור האפשריות בכתב הדבאנאגארי. לכן פותח גופן המאפשר ליצור חלק מצורות הקישור באמצעות הרכבה של גלופות.

כדי לסדר דפוס ב־TEX באופן לא־לינארי יש להשתמש בפקודות מורכבות. לכן הקלט הדרוש ל־TEX לסידור דפוס דבאנאגארי הוא לא במיוחד קריא.

devnag היא תוכנת C המשמשת **מעבד מקדים** (preprocessor) ל־TEX. היא מקבלת קלט בתעתיק לטיני והופכת אותו לאוסף פקודות TEX.

```
foo.dn   {\dn devanAgarI hindI}
foo.tex  {\dn d\?vnAgrF Eh\306wdF}
foo.pdf  देवनागरी हिन्दी
```

dev היא חבילת LATEX שמסדרת את הפלט של devnag. אין תמיכה בקידודים הודיים. החבילה תואמת babel, אך לא משתמשת באותו הממשק למעבר שפה.

קיימות חבילות לכתבים הודיים נוספים, הפועלות על עקרונות דומים.

הכתב הערבי

בכתב הערבי יש לכל אות ארבע צורות בסיסיות, בהתאם למקומה במילה.

תחילית	אמצעית	סופית	נפרדת
ء	ء	ء	ء
ع	ع	ع	ع
ق	ق	ق	ق
ه	ه	ه	ه

בנוסף, קיימות צורות חיבור רבות.

شجرة ← شجرة في ← في تمر ← تمر بها ← بها

نبت ← نبت قبر ← قبر بكم ← بكم لحم ← لحم

הכתיבה מימין לשמאל; ספרות נכתבות משמאל לימין; תנועות נכתבות מעל ומתחת לעיצורים.

ArabTEX (פיתוח: Klaus Lagally)

מספר הגלופות הנדרשות על מנת להכיל את כל צורות הקישור גדול בהרבה מ־256. אבל אותיות רבות בנויות על בסיס זהה, ונבדלות אך ורק בנקודות (למשל ج, ح, و־خ). לכן יש **פסאודו־גופן** עם צורות בסיס המורכבות ע"י TEX.

הקלט הוא בתעתיק לטיני; ניתן גם להשתמש בקידודים ערביים, כולל UTF-8. המעבד המקדים כתוב ב־TEX, ולכן העיבוד מתבצע בריצה אחת.

```
foo.tex  <muntadY lInuks - .haifA>
foo.pdf  مُنتَدَى لِينُكس - حَيفَا
```

תמיכה בשפות שונות (ערבית, פרסית, אורדו, טורקית עותומאנית, מלאית עתיקה ועוד). תמיכה בפלט בתעתיק.

תמיכה דו־כיוונית משתמעת: רצף תווים הפותח בספרה (אחרי רווח) נכתב משמאל לימין, כל רצף אחר נכתב מימין לשמאל.

תואמת babel, אבל לא משתמשת באותו הממשק למעבר שפה.

הכתב הקוריאני

בכתב הקוריאני (האנגוּל 한글) האותיות מסודרות כך שכל הברה תיצור ריבוע.
עיצור שבא לפני **תנועה** נכתב מעליה ו/או לשמאלה, תלוי בצורה של סימן התנועה. (אם ההברה פותחת בתנועה יבוא לפניה עיצור־דמה ㅇ).

ע"ת **CV**

ㅊ ch^h + ㅣ i = 치 ch^hi

ㅅ s + ㅜ u = 수 su ㅎ h + ㅚ wi = 회 hwi

אם בסוף ההברה יש **עיצור** הוא נכתב מתחת ל**תנועה** ול**עיצור** הראשון.

עת"ע **CVC**

ㄱ k + ㅣ i + ㅁ m = 김 kim

שני עיצורים בסוף ההברה נכתבים מ**שמאל** ל**ימין**.

עתע"ע **CVCC**

ㄷ t + ㅏ a + ㄹ l + ㄱ k = 닭 $talk$

ההברות מצטרפות משמאל לימין, או מלמעלה למטה.

서 울 $seoul$ 서울 = (ㅅ s + ㅓ eo) + (ㅇ + ㅜ u + ㄹ l)

TEX קוריאני

קיימות שתי חבילות עיקריות: אחת שפותחה בקוריאה, והאחרת CJK של Werner Lemberg התומכת גם בסינית וביפאנית.

מתייחסים לכל הברה מורכבת כאל תו נפרד. הדבר דורש שימוש במספר רב של גופנים בני 256 תווים, ובבחירה אוטומטית של הגופן לפי התו הנדרש.

הקלט בקידודים קוריאניים (או סיניים ויפאניים), אין תמיכה בקלט בתעתיק.

גופני TEX כיום מכילים את כל ההברות הדרושות לסידור קוריאנית מודרנית, אך לא את ההברות הדרושות לסידור קוריאנית של ימי הביניים.

סידור דפוס עברי

דרישות:

- קלט מתאים
- פלט (גופנים)
- ניתוח קונטקסטואלי (?)
- תמיכה בניקוד ובטעמים
- תמיכה דו־כיוונית

תקווה Tiqwah (פיתוח: Γιάννης Χαραλάμπους)

מיועדת לסידור טקסט מקראי.

לא חופשית, לא נגישה לציבור הרחב.

קלט בתעתיק לטיני־מדעי. מעבד מקדים כתוב ב־GNU Flex. שימוש ב־TEX סטנדרטי. ניתוח קונטקסטואלי של אותיות סופיות.

גופנים "מקראיים", בהשראת ה־BHS.

תמיכה בניקוד (טברני, בבלי וארץ־ישראלי) ובטעמים. אלגוריתם מורכב קובע את מקום הניקוד והטעם ביחס לאות, ועשוי להשפיע גם על ריווח האותיות במילה.

תמיכה בתופעות דפוס ייחודיות למקרא (אותיות תלויות, הפוכות, שבורות וכד').

אין תמיכה בספרות ובסימני פיסוק מודרניים.

מקור Makor (פיתוח: Alan Hoenig)

חופשית.

קלט בתעתיק לטיני אשכנזי־אמריקאי. דורש הרחבה דו־כיוונית של TEX. ניתוח קונטקסטואלי של אותיות סופיות.

מספר גופנים שמישים. מנגנון הגדרת גופנים לא סטנדרטי. `\catcode` לא סטנדרטיים. מנגנון החלפת שפה אינו תואם babel.

תמיכה אוטומטית בניקוד. אין תמיכה בטעמים.

HebTEX (פיתוח: Klaus Lagally)

מוד של ArabTEX. לא עובד יחד עם המוד הערבי.

קלט בתעתיק לטיני־מדעי. שימוש ב־TEX סטנדרטי. ניתוח קונטקסטואלי של אותיות סופיות.

מספר גופנים מצומצם.

תמיכה אוטומטית בניקוד. אין תמיכה בטעמים. תמיכה בפלט בתעתיק.

דו־כיווניות משתמעת: רצף תווים הפותח בספרה (אחרי רווח) נכתב משמאל לימין, כל רצף אחר נכתב מימין לשמאל.

תמיכה ראשונית בערבית־יהודית.

cjhebrew (פיתוח: Christian Justen)

מיועדת לקטעים עבריים קצרים בתוך טקסט לטיני.

קלט בתעתיק לטיני־מדעי. שימוש ב־TEX סטנדרטי. ניתוח קונטקסטואלי של אותיות סופיות.

גופנים ייחודיים.

תמיכה אוטומטית בניקוד. אין תמיכה בטעמים. אין תמיכה בספרות ובסימני פיסוק מודרניים.

Λ/Ω (פיתוח: Γιάννης Χαραλάμπους ו־John Plaice)

מערכת רב־לשונית מבוססת על יוניקוד. תומכת גם בעברית.

המערכת הישראלית HebLaTeX

פיתוח מסוף שנות ה־80 בטכניון/אונ' עברית (?). חבילה עברית ל־LaTeX 2.09 (טרום 1994). בסביבות 1997 פיצול: נדב הראל ממשיך את החבילה של 2.09 (מיועדת בעיקר לשימושו הפרטי), בוריס Lavva (איך כותבים את שמו בעברית?) מתאים את החבילה ל־LaTeX 2ε ומשלב אותה ב־babel.

החבילה דורשת הרחבה דו־כיוונית של TeX.

קידוד קלט

iso-8859-8, cp-1255 (אותיות בעברית בטווח 0xE0–0xFA) וכן cp-862 (אותיות בעברית בטווח 0x80–0x9A) מטופלים ע"י החבילה inputenc: תווי הקלט הגבוהים הם אקטיביים, ומומרים בפקודות כגון \hebalef.

משמעות: לא ניתן להגדיר פקודות בנות יותר מתו עברי אחד, כגון \הדגש.

si-960 (אותיות בעברית בטווח 0x60–0x7A) דורש קידוד פלט מתאים.

קידוד פלט

קידוד בן 7 ביטים LHE מכיל אותיות, ספרות וסימני פיסוק;
קידוד בן 8 ביטים HE8 מכיל גם ניקוד וטעמים (עדיין בפיתוח).

הגופנים בקידוד LHE הם ישנים ולא מתבצע בהם פיתוח. הכוונה היא לנטוש את הקידוד הזה. נחוץ בעיקר למסמכים שמסתמכים על הגופנים הישנים, ולמסמכים ישנים שכתובים בקוד si-960.

גופני קולמוס החדשים (פיתוח: מקסים יורש) מקודדים ב־HE8. קידודים דומים פותחו באונ׳ ת״א עבור גופנים של מיקרוסופט ו־IBM.

בחירת קידוד הגופן:

```
\def\HeblatexEncoding{HE8}
\def\HeblatexEncodingFile{he8enc}
```

בגירסאות החדשות, HE8 היא ברירת המחדל.

תמיכה בניקוד ובטעמים

גופני קולמוס כוללים סימני ניקוד וטעמים, אבל יש לקבוע את מיקומם ביחס לאות באופן ידני. מנגנון למיקום אוטומטי של סימני ניקוד נמצא בפיתוח.

קיצורים לאותיות מודגשות ממומשים ברמת הגופן, לא ברמת babel.

תמיכה דו־כיוונית

אין תמיכה משתמעת. בחירת השפה קובעת את הכיוון. למעבר שפה בין פסקאות יש להשתמש בפקודות הרגילות של babel. למעבר שפה בתוך פיסקה קיימות הפקודות `\L{}` (טקסט לטיני בפיסקה עברית) ו־`\R{}` (טקסט עברי בפיסקה לטינית).

אין פקודה אחידה למעבר כיוון בלי לשנות שפה (למשל לצורך כתיבת מספר). ניתן להגדיר פקודה כזו, למשל:

```
\newcommand{\N}[1]{\beginL#1\endL}
```

אין שיקוף של סימנים מתחלפים כגון סוגריים.

hebcal

חבילה נוספת שמחשבת תאריכים עבריים.

ולסיום, מערכת חדשה ומבטיחה:

TAUX (פיתוח: ענת רפופורט וסיון טולדו)

מערכת מבוססת־יוניקוד. תואמת לסטנדרטים, כולל הסטנדרט הדו־כיווני. בפיתוח (אלפא).