סִדּוּר דְּפוּס רַב־לְשׁוֹנִי

Multilingual typesetting أَلتَّنْضِيدُ آلْمُتَعَدِّدُ آللُّغَاتِ

Ron Artstein רון ארטשטיין رون ارتشتين

Technion טכניון تخنيون

artstein@cs.technion.ac.il

29 בספטמבר 2003 ג׳ בתשרי, תשס״ד

מהו סידור דפוס

- פותח באירופה במאה ה־15.
- גלופות של אותיות בודדות מסודרות בשורות, אשר מסודרות בעמודים.
- מהיר וזול בהרבה מחריטת גלופות של עמודים שלמים.
- מאפשר הדפסה במספר רב של עותקים.

עד היום לא פותח תהליך מכני או חישובי המסוגל להגיע לאיכות של כתב־יד הנכתב בידי אומן. כאשר נדרשת איכות יוצאת דופן ממשיכים להשתמש בשיטה האיטית והיקרה של העתקת כתבי־יד (למשל: ספרי תורה).

עם השנים התפתחה מגמה של ייעול והוזלת תהליך הדפוס, על חשבון איכות הדף המודפס. המעבר לסידור דפוס ממוחשב ממשיך מגמה זו.

ציפיות מסדר־דפוס ממוחשב

סְדַר־דפוס: מכונה המשמשת לסידור הגלופות. המקבילה בעולם המחשבים: שפת מדפסות (דוגמת PostScript או PCL).
סַדָּר־דפוס: אומן, אשר מכין את הגלופות על פי כתב היד והוראות המהדיר. זקוק לידע מקצועי רב לצורך מלאכתו.

אנו מצפים מתוכנת סדר־דפוס שתבצע חלק מעבודתו של הסַדָּר האנושי; לכן צריך לבנות ידע רב לתוך התוכנה.

הדרישה העיקרית מסדר־דפוס ממוחשב

- לתת לכותב/עורך/מהדיר **שליטה מלאה** לגבי מיקום התווים על הדף המודפס.

(הערה: שליטה מלאה איננה ערובה לתוצאה מוצלחת, כי כותבים ועורכים רבים אינם בעלי הידע הנדרש בתחום סידור־הדפוס.)

דרישות נוספות:

- עריכה נוחה של הטקסט ואפשרות לעריכה חוזרת.
- אוטומציה של עבודת הסַדָּר: חלוקת הטקסט לשורות, קביעת מרווחים וגודל התווים, ארגון הטקסט בעמוד.
- אוטומציה של עבודת המהדיר: קביעת גופנים, כותרות, עיצוב; וכן סידור אוטומטי של מראי מקומות, הערות שוליים, תוכן עניינים, רשימת מקורות, אינדקס, מעברי עמודים ועוד.
- גמישות: שימוש חוזר בתבניות קבועות ושמירת טקסט אלטרנטיבי.
- יכולת להתממשק עם תוכנות חיצוניות (מאייתים, עיבוד שפה ועוד).
- קבלה וסידור של קלט מתוכניות חיצוניות (תרשימים, תמונות).
- יצירת פלט מסוגים שונים (טקסט מסומן, קבצי נתונים, שמע).

הרצאה זו תתמקד בדרישות שנחוצות לסידור רב־לשוני.

קצת על עיבוד תמלילים

עיבוד תמלילים הוא דרך עבודה מול מחשב עם המאפיינים הבאים:

1. המטרה היא בעיקר עריכה של טקסט להדפסה.
2. העבודה מתבצעת מול תוכנה אחת בממשק אחד.
3. הגישה לקובץ מתבצעת דרך הממשק, ומבנה הנתונים שקוף למשתמש.

מעבדי תמלילים רבים מתאפיינים גם בתכונות הבאות:

4. תצוגת הקובץ על המסך דומה לפלט המודפס (WYSIWYG).
5. התצוגה מתעדכנת עם כל לחיצת מקש.

באופן עקרוני אין מניעה להשתמש בדרך עבודה זו לסידור דפוס, אבל **תכונה 3** מקשה מאוד על קבלת **שליטה מלאה** בתוצאה הסופית, ולכן תוכנות עיבוד תמלילים בדרך כלל אינן מסוגלות לסדר דפוס באיכות גבוהה.

תוכנות סידור דפוס

- **TEX:** 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.
- **יציבה מאוד:** גירסה 3.14159 ב־1995, 3.141592 ב־2002 (תיקוני באגים בלבד). מספר הגירסה מתכנס ל־π, כל שינוי מוסיף ספרה עשרונית. 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, או פורמט אחר בגירסאות שונות.
- שפת תיכנות. ניתנת לקונפיגורציה ולהרחבה. כמעט לא עובדים מול TEX ישירות. הרחבות רבות, ברמות שונות של נגישות, אמינות ותיעוד.

תוכנות אחרות: קניניות, בשימוש בתי־דפוס, לא מוכרות בציבור הרחב.

סדר האלף־בית

נחוץ לסידור אוטומטי של רשימות. משתנה משפה לשפה.

בספרדית מתייחסים לצירוף **ch** כאל אות נפרדת, הבאה אחרי **c**.

en**c**umbrar, en**c**urtir, en**ch**apar

בגרמנית מקובל כיום להתייחס לאות **ä** כשקולה לאות **a**; בעבר היה נהוג להתייחס אליה כשקולה לצירוף **ae**.

B**a**d, B**a**hn, B**ä**r, B**a**um
B**ad**, B**ä**r, B**ah**n, B**au**m

בשבדית האותיות **å**, **ä**, **ö** באות בסוף האלף־בית, אחרי **z**.

בהונגרית מתייחסים לצירוף **cs** כאל אות נפרדת הבאה אחרי **c**, ולאותיות **a** ו־**á** כשקולות.

cukor, **c**uppant, **cs**alit, **cs**ata
alma, **á**lom, **a**lorvos

מספור אוטומטי

סדר־דפוס צריך לעתים להכניס מספור אוטומטי ברשימות. בשפות רבות המשתמשות בכתב הלטיני מקובל להשתמש באותיות לפי סדר האלף־בית. מה שנחשב ל"אות" לצורך העניין איננו בהכרח זהה למה שנחשב ל"אות" לצורך סידור מילונאי.

ספרדית: ABCDEFGHIJKLMN**Ñ**OPQRSTUVWXYZ

שבדית: ABCDEFGHIJKLMNOPQRSTUVWXYZ**ÅÄÖ**

בכתב העברי אין משתמשים בסדר האלף־בית, אלא בשיטת מספור מסורתית. כך גם בכתב היווני והערבי.

א, ב, ג, . . . , י, יא, . . . , טו, טז, . . .

מיקוף אוטומטי

על מנת שהטקסט יהיה קל לקריאה, רצוי שהשוליים יהיו ישרים, ושהרווחים בין המילים יהיו שווים ככל האפשר. לשם כך יש צורך לעתים לחלק מילה לשניים.

המיקוף הנכון נקבע בין היתר על־פי השפה.

אנגלית sig-nal [ˈsɪg.nl̩]

צרפתית si-gnal [siˈɲal]

בשפות מסוימות, מיקוף נכון הוא יותר מאשר מציאת מקום הולם לחלוקה.

בגרמנית, כאשר העיצור **k** מוכפל הוא נכתב **ck**, אבל אם העיצור הכפול מתחלק על פני שתי שורות הוא נכתב **k-k**.

Be**ck**er, Be**k-k**er

ב**גרמנית**, במילה המורכבת משני חלקים שהראשון מסתיים בעיצור כפול והשני באותו עיצור, נכתב העיצור בשני תווים; בחלוקה לשתי שורות נכתב העיצור בשלושה תווים.

Schiff + Fahrt = Schi**ff**ahrt, Schi**ff-f**ahrt

ב**הונגרית**, הכפלת עיצור הנכתב כצירוף של שני תווים מסומנת על־ידי הכפלת התו הראשון; בחלוקה לשתי שורות, על־ידי הכפלת הצירוף כולו.

ga**ll**yak, ga**ly-ly**ak

hos**sz**ú, hos**z-sz**ú

villamos + szék = villamo**ssz**ék, villamo**s-sz**ék :אבל

גורמים נוספים שקובעים את המיקוף הנכון:

הגייה: **אנגלית**		ומבנה: **גרמנית**	
rec-ord	['rɛk.ɻd]	wach-stube	חדר־משמר
re-cord	[rɪ'kɔrd]	wachs-tube	שפופרת שעווה

מסורות דפוס שונות

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 ?

(מסורות אינן בהכרח לפי שפה: יש מסורות לפי ארץ, קבוצה אתנית וכד'.)

איך לסמן את השפה

כתב־יד שנמסר לסידור מכיל את הטקסט עצמו, וכן הוראות של המהדיר. כך גם בקובץ שנקרא ע"י סדר ממוחשב, למשל קובץ SGML: הוראות בכחול, טקסט בירוק.

```
<marginnote> Some text </marginnote>
```

ההבחנה בין טקסט להוראות איננה תמיד חדה. מבחינת הכותב, התווים בשחור בקובץ הבא הם טקסט; מבחינת התוכנה LaTeX 2ε, התו à הוא דווקא הוראה.

```
\documentclass{article}
\usepackage[latin1]{inputenc}
\begin{document}
voilà!
\end{document}
```

שאלה: איך צריך לקודד אינפורמציה לגבי שפה? בתווים או בהוראות?

קידוד בתווים ובסימון החיצוני

האינפורמציה לגבי התוצאה הסופית הרצויה מתחלקת בין התווים להוראות. החלוקה יכולה להתבצע באופנים שונים. נניח שאנו רוצים לסדר את הפלט הבא:

George *Washington*

בקידודים סטנדרטיים האותיות הרישיות והקטנות מהוות תווים נפרדים, ואילו ההבחנה בין אותיות ישרות לנטויות נעשית באמצעות הוראות.

George <italic> Washington </italic>

אבל באופן עקרוני אפשר גם להיפך.

<uppercase>g</uppercase>eorge <uppercase>*w*</uppercase>*ashington*

מהו תו?

התפישה של Unicode: קוד נפרד לכל תו. אבל לא ברור בדיוק מהו תו.

רומנית משתמשים באותיות ş ו־ţ עם פסיק למטה (מבוטאות כמו 'ש' ו־'צ').
מדפיסים מסוימים החליפו את הפסיק ב־cedilla, ככל הנראה מחוסר ידע.

Ţ~ţ [ts] Ţ~ţ Ş~ş [ʃ] Ş~ş

(האות ş נמצאת בשימוש בטורקית; האות ţ איננה בשימוש באף שפה.)

האם ş ו־ş הם תו אחד או שני תווים שונים?
אם הם תו אחד, אז צריך לקבוע את התצוגה בפלט לפי השפה.

The Unicode Standard 4.0 (online edition), Section 7.1, pages 168–169.
http://www.unicode.org/versions/Unicode4.0.0/ch07.pdf

In Turkish and Romanian, a cedilla and a comma below sometimes replace one another depending on the font style. However, the form with cedilla is preferred in Turkish, and the form with comma below is preferred in Romanian. The characters with explicit commas below are provided to permit the distinction from characters with cedilla. However, legacy encodings for these characters contain only a single form of each of these characters. ISO/IEC 8859-2 maps these to the form with cedilla, while ISO/IEC 8859-16 maps them to the form with comma below. Migrating Romanian 8-bit data to Unicode should be done with care.

טבלאות המרה לאותיות רישיות

הבחנה בין אותיות רישיות לקטנות קיימת בכתב הלטיני (A~a ,B~b), בכתב הקירילי (А~а ,Б~б) ובכתב היווני (Α~α ,Β~β). סדר־דפוס נדרש לעתים לבצע המרה מאותיות רישיות לקטנות או להיפך.

טבלת המרה חד־חד ערכית איננה אפשרית:

- אות רישית אחת יכולה להיות המקבילה של אותיות קטנות שונות: באיסלנדית קיים הזוג Ð~ð, ובקרואטית קיים הזוג Đ~đ.
- זוג אותיות שמתקיים בשפה אחת עשוי שלא להתקיים בשפה אחרת, גם אם האותיות עצמן קיימות: במרבית השפות הנכתבות בכתב הלטיני מתקיימת ההמרה I~i, אבל בטורקית ההמרה היא דווקא İ~i ו־I~ı.
- אות קטנה יכולה להיות מקבילה לצירוף של אותיות רישיות: בגרמנית מתקיימת ההמרה ß~SS (וכן ss~SS).

פתרונות: קידוד נפרד לאותיות זהות בעלות המרה שונה; סימון חיצוני של השפה; התעלמות מהבעיה.

מרכאות פותחות וסוגרות

אנגלית	“some text”
צרפתית	« un texte »
גרמנית	»ein Text« „ein Text“
ערבית	‹‹نصّ››
עברית	”טקסט“ „טקסט”

אפשרויות לייצוג:
— תו נפרד לכל סוג של מרכאה בקלט;
— תו גנרי לפתיחה/סגירה, המרה לתו פלט מתאים לפי השפה;
— תו גנרי לפתיחה/סגירה, בחירת גופן לפי השפה.

אין כיום קידוד סטנדרטי המבחין בין מרכאות פותחות וסוגרות בעברית.

ניתוח קונטקסטואלי (ערבית)

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

נפרדת	סופית	אמצעית	תחילית
ع	ـع	ـعـ	عـ
ق	ـق	ـقـ	قـ
ه	ـه	ـهـ	هـ

בנוסף, קיימות צורות חיבור רבות.

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

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

הסדר צריך לבחור את הגלופה המתאימה לפי ההקשר.

ניתוח קונטקסטואלי (עברית)

בכתב העברי יש חמש אותיות בעלות שתי צורות בסיסיות, בהתאם למקומן במילה.

תחילית: כ מ נ פ צ

סופית: ך ם ן ף ץ

מקובל שלא לבצע ניתוח קונטקסטואלי, אלא לקודד כל צורה כתו נפרד.

סיבה היסטורית: במכונת כתיבה מכאנית קל יותר לתת תו נפרד לכל צורה.
סיבה לשונית: במקרים רבים ניתוח קונטקסטואלי לא יעבוד: ג'יפ, תנ"ך, בד"כ.

"אותיות מנצפ"ך בסוף ראשי תיבות נכתבות כסופיות אם ראשי התיבות מבוטאים כמילה אחת (**או"ם, תָּנָ"ך**), אך אם התיבות נקראות במלואן או אם קוראים את האותיות בשמן, נכתבות הן כלא־סופיות (**אח"כ, חוה"מ, מ"מ** [ממלא מקום], **מ"מ** [מֵם מֵם])."
(לובה חרל"פ, הפיסוק — בין תחביר לסגנון, **בלשנות עברית 52**, ניסן תשס"ג/אפריל 2003, 37–51; הערה 30, עמוד 50).

הכותבת מצטטת ממקור אחר (חוברת "לשוננו לעם"), ולא ברור על מה מסתמך חיבור זה. התרשמותי האישית היא שהתיאור לעיל איננו תמיד משקף את השימוש הרווח בעברית: **אב"כ, בע"מ** נהגות כמילים אבל נכתבות באותיות לא־סופיות; כמו־כן צירופים רבים ניתנים לכתיבה בשני אופנים: **קב"ן/קב"נ, פק"ס/פק"מ, קמב"ץ/קמב"צ.**

ניתוח קונטקסטואלי (עברית)

בכתב העברי יש חמש אותיות בעלות שתי צורות בסיסיות, בהתאם למקומן במילה.

תחילית: כ מ נ פ צ

סופית: ך ם ן ף ץ

מקובל שלא לבצע ניתוח קונטקסטואלי, אלא לקודד כל צורה כתו נפרד.

The Unicode Standard 4.0 (online edition), Section 8.1, page 193.
http://www.unicode.org/versions/Unicode4.0.0/ch08.pdf

Final (Contextual Variant) Letterforms. Variant forms of five Hebrew letters are encoded as separate characters in this block, as in Hebrew standards including ISO/IEC 8859-8. These variant forms are generally used in place of the nominal letterforms at the end of words. Certain words, however, are spelled with nominal rather than final forms, particularly names and foreign borrowings in Hebrew, and some words in Yiddish. Because final form usage is a matter of spelling convention, software should not automatically substitute final forms for nominal forms at the end of words. The positional variants should be coded directly and rendered one-to-one via their own glyphs—that is, without contextual analysis.

ניתוח קונטקסטואלי (גרמנית)

בכתב הגרמני שהיה בשימוש עד 1941 התקיימה הבחנה בין שתי צורות של האות s: ſ בתחילת או באמצע הברה, ß בסוף הברה.

ככלל, נחוץ ניתוח מאוד מתוחכם כדי לזהות גבולות של הברות בטקסט רציף, וזה לא תמיד אפשרי.

$$\mathfrak{Wach} + \mathfrak{Stube} = \mathfrak{Wach\text{ſ}tube}$$
$$\mathfrak{Wachs} + \mathfrak{Tube} = \mathfrak{Wach\text{ß}tube}$$

לכן נחוץ קידוד נפרד לכל צורה של s. קידוד כזה קיים ב־Unicode, אבל לא במרבית לוחות המקשים...

צורות חיבור אסתטיות 1

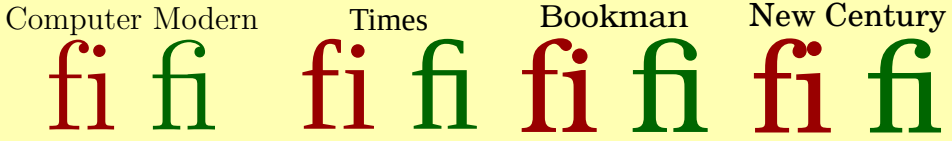
בגופנים רבים, החלק העליון של האות f חורג מרוחב האות; הדבר מביא לתוצאה לא אסתטית בסמיכות לאות i. לכן קיימות בגופנים רבים צורות חיבור מיוחדות. TEX מבצעת את ההחלפה באופן אוטומטי.

בעייה. טורקית משתמשת בשתי האותיות i ו־ı. מה לגבי צורת החיבור fi? האם יש לקרוא אותה כ־f+i או כ־f+ı?

פיתרון. כשכותבים בטורקית, אין להשתמש בצורות החיבור; לקבלת תוצאה אסתטית, רצוי להשתמש בגופן שבו החלק העליון של f אינו חורג מרוחב האות, למשל Palatino: fıstık 'פיסטוק', fistan 'שמלה'.

צורות חיבור אסתטיות 2

בגופנים רבים, החלק העליון של האות f חורג מרוחב האות; הדבר מביא לתוצאה לא אסתטית בסמיכות לאות i. לכן קיימות בגופנים רבים צורות חיבור מיוחדות. TEX מבצעת את ההחלפה באופן אוטומטי.

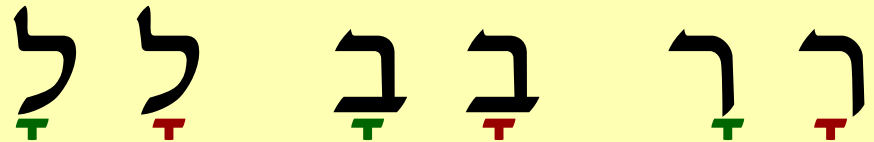


בעייה. צירוף האותיות fj נפוץ בנורבגית, ומביא לבעיות דומות לאלו של fi. אבל גופנים רבים נכתבו עבור אנגלית, ואינם כוללים צורות חיבור עבור fj.

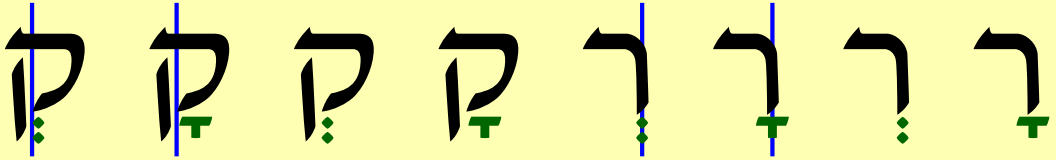
פיתרון. כשכותבים בנורבגית, רצוי להשתמש בגופנים מיוחדים שיש בהם את צורת החיבור fj, או בגופן שבו החלק העליון של f אינו חורג מרוחב האות.

ניקוד 1

לכל אות בעברית יש ציר אורך מסויים, שמתחתיו יש לשים את הניקוד.



באותיות מסוימות ציר האורך מתייחס למרכז סימן הניקוד, ואילו באחרות הוא מתייחס לאחד הקצוות.



ניקוד 2

גם לניקוד העליון צירי אורך משלו.

רֵ מֵ לֵ

והיחס בין החולם לאות וי"ו תלוי בשימוש שלה, כעיצור או כאם קריאה.

עָוֹן אוֹר

אין כיום קידוד סטנדרטי המבחין בין וי"ו כעיצור וכאם קריאה.

רב־כיווניות

בעת סידור כתב רב־כיווני יש לדעת את הסדר הלוגי של הכתיבה, למקרה שמעבר שורה עשוי להתרחש בקטע בו כיוון הכתיבה שונה מהכיוון הראשי.

מטוס חדיש מדגם בואינג 757־ 300 נכנס היום לשירות.

מטוס חדיש מדגם בואינג 300־ 757 נכנס היום לשירות.

סדר אנושי מקבל כתב־יד שבו התווים מופיעים בסדר ויזואלי, ומסיק את הסדר הלוגי על פי ידיעותיו ונסיונו.

לסדר ממוחשב יותר נוח להזין את התווים בסדר לוגי, ולקבוע כללים לתצוגה רב־כיוונית.

- סימון ישיר של כיוון הכתיבה באופן חיצוני (תפישת LaTeX/Babel)
- קביעת כיוון הכתיבה באופן משתמע (תפישת Unicode, ArabTeX)

קביעה משתמעת

קביעה משתמעת איננה תמיד חד־משמעית: איך יש להציג את הספרות במחרוזת הבאה (בסדר לוגי מימין לשמאל)? תלוי בכוונת הכותב.

מ ע ב ד י ␣ I n t e l ␣ 8 0 8 6 ␣ נ ח ש ב י ם

מעבדי Intel 8086 נחשבים מעבדי Intel 8086 נחשבים

הנה דוגמאות שמראות שאכן שני הסדרים נדרשים; הסדר הלוגי של התווים זהה בשני המקרים — רצף עברי, רווח, רצף לטיני, רווח, רצף ספרות, רווח, רצף עברי.

מעבד Intel 386 מיושן מעבד Intel 900 מה"ץ חדיש

לכן יש לקבוע בכל מנגנון של קביעת כיוון משתמעת אפשרויות לשליטה ישירה בכיווניות. ב־Unicode הדבר מתבצע ע"י הכנסת תווים מיוחדים לקביעת הכיוון, שאינם מופיעים בפלט המודפס.

המקף בעברית

בסימון טווח של מספרים במקף מפריד, יש הכותבים את המספרים מימין לשמאל (זוהי הנחיית האקדמיה), ויש הכותבים משמאל לימין.

עמ' 35–37 עמ' 37–35

אם משמיטים חלק מהספרות, הטווח נכתב תמיד משמאל לימין.

בשנים 1987–1992 בשנים 1987–92

הדבר יכול להביא לדו־משמעות בטקסט המודפס: האם הפלט עמ' 87–121 מתכוון לטווח מ־87 עד 121, או מ־121 עד 187?

כיווניות תלויית כתב ושפה

הצגה משתמעת של טקסט דו־כיווני חייבת להתחשב בכתב: סימן האחוז בא מימין למספר בכתב העברי, ומשמאלו בכתב הערבי.

אנגלית 4.5% fat

עברית 4.5% שומן

ערבית ٤٫٥% دهن

ההצגה תלויה גם בשפה.

כך כותבים מיליארד בערבית: ١٠^٩ ובפרסית: ۱۰^۹

כיווניות תלוית־משמעות

הצגת המחרוזת "1, לוכסן, 4" תלויה לא רק בכתב ובשפה אלא גם במשמעות.

	"רבע"	"אחד באפריל"
אנגלית בריטית	1/4	1/4
עברית	**1/4**	**1/4**
ערבית	١/٤	١/٤

איך ידע סדר־הדפוס לקבוע את הכיוון?
— תווים שונים עבור משמעויות שונות של הלוכסן;
— הוראות כיוון מפורשות;
— סימון חיצוני של המשמעות.

נ.ב. בפרסית, משמעות הייצוג הויזואלי ١/٤ היא "ארבעה בינואר" או "אחת וארבע עשיריות" (שבר עשרוני); המשמעות "רבע" נכתבת ١÷٤.

סימנים מתחלפים

סימנים רבים באים בזוגות: () [] {} ועוד.

- הצגה קבועה: קוד 0x28 מייצג סוגר שמאלי, 0x29 מייצג סוגר ימני.
  (תפישת LaTeX/Babel)

  דורש הכנסת הקוד המתאים בהתאם לכיוון בטקסט מיוצר אוטומטית.
- הצגה מתחלפת: קוד 0x28 מייצג סוגר פותח, 0x29 מייצג סוגר סוגר.
  (תפישת Unicode, ArabTeX)

  דורש ייצוג לפי הקשר בתוכנות: עורכים, סדר הדפוס, תוכניות הצגה.

גם סימנים מתמטיים עשויים להיות מתחלפים בטקסט דו־כיווני:

עבור כל $3 \leq i$ לא קיימים a, b, c טבעיים כך ש: $a^i + b^i = c^i$

עורך הטקסט

- תצוגה נוחה לקריאה (כולל תצוגה דו־כיוונית).
- תצוגה נוחה לעריכה (רצוי חד־מימדית, ללא דו־משמעות).
- סדר מתוחכם מבצע יותר מדי חישובים מכדי שהעורך יוכל להשתמש בסדר כדי לרענן את התצוגה (מסמך זה מתקמפל תוך כ־3 שניות).

  מאמר של Jonathan Fine ב־TUGboat 22(4), 2001 (יצא לאור ביוני 2003) מציג אפשרות להריץ את TEX כ־daemon שמקבל קטעים קצרים של קובץ מקור לסידור, באופן שמאפשר להשתמש בסדר עצמו לרענון התצוגה; המערכת עדיין בפיתוח.

פלט לשימוש חיצוני

- קבצי המקור המוזנים לסדר משמשים לא רק לסידור דפוס, אלא גם לפלט מסוג אחר (למשל טקסט מסומן לוגית כגון html).
- הפלט חייב להתאים לסטנדרטים.
- אם הסדר מתוחכם מאוד, רק הסדר עצמו יכול להבין את הקלט שלו.

המרה מכתב לכתב

שפת **האוסה** (ניגריה, כ־24 מיליון דוברים) נכתבת בשני כתבים: הלטיני והערבי.

הכתב הלטיני מבחין בין חמש התנועות, אך לא בין תנועות קצרות לארוכות.
הכתב הערבי מבחין בין תנועות קצרות לארוכות, אך לא בין התנועות o ו־u.
(אף אחד מהכתבים אינו מציין טונים או מבחין בין r ל־ɾ.)

ייצוג פונטי	כתב לטיני	כתב ערבי	משמעות
[duːkàː]	duka	دُوكَا	'הכאה'
[dukà]	duka	دُكَ	'כל'
[sûː]	su	سُو	'דיג'
[sôː]	so	سُو	'תשוקה'

(מקור: http://www.humnet.ucla.edu/humnet/aflang/Hausa/Pronunciation/writing.html)

כדי ליצור פלט בשני הכתבים **מקובץ מקור אחד** יש לבחור ייצוג הכולל את כל ההבחנות, ולבצע המרות מתאימות.