

אות לאות, קו לקו:

תוכנת סריקה וזיהוי אותיות עברית ולועזית

(מאמר שישי בסידרה)

מאת משה פלאי

מעבד מגירסת 80386 ומעלה. נדרש זיכרון פנימי של 3MB 'עכבר', וכן מערכת הפעלה של חלונות מגירסת 3.0 ומעלה. בנוסף על כך נדרש סורק אופטי, שהוא מכשיר נפרד דמוי מכונת העתקה, המתחבר למחשב, כסורק היולט פקד, זירוקס, או דומיהם, מתוך שורה של סורקים ידועים. בית התוכנה אינו ממליץ על שימוש בסורק-יד, מאלו המצויים בשוק, שהם אמנם זולים יותר, אך בלתי יעילים. אי אפשר להפעיל את התוכנה עצמה בלי החומרה הנזכרת, דהיינו: בלי סורק, ובלי מערכת ההפעלה של חלונות. אמנם בעבר הכינה החברה גרסה למערכת ההפעלה של דוס, אך עם השימוש הגובר בחלונות, נטשה אותה והמשיכה להתפתח בכיוון החלונות הפתוחים. התוכנה ליגטורה באה עם תקע-הגנה, כרוב התוכנות הישראליות, המוגנות מפני התופעה שבענה המחשבת קרויה 'גונבה'... (על משקל: חומרה, תוכנה), דהיינו: תוכנה בלתי-חוקית. התוכנה לא תפעל, איפוא, בלא מתקן ההגנה. יש לציין, כי באחרונה היינו כמה בתי התוכנה להפיק תוכנות בלתי מוגנות מתוך ביטחון ביושרו של הלקוח העברי.

כיצד פועלת תוכנת הסורק העברית?

התקנת התוכנה תעשה לפי ההוראות המצויות במדריך. הקש את הפקודה 'INSTALL' והמערכת תותקן אוטומטית תוך מתן אפשרות לבחירת ברירת-מחדל. מלבד המדריך, תוכל להיוועץ בקובץ הנקרא README.TXT. עליך להתקין את המערכת בקונפיגורציה הנדרשת לך לפי הסורק העומד לרשותך.

לאחר התקנת התוכנה והטענתה, כמקובל בחלונות, עליך לקבוע את הנתונים של עבודתך. בשלב הראשון חר בכפתור ההפעלה הראשון הנקרא INPUT--קלט--, וכאן עליך להזין מידע לתוכנה ולציין מהו מקור הסריקה: אם רצונך להתחיל להשתמש בסורק, או לטעון קבצים מוכנים שכבר נסרקו בעבר, אפשרות שתתרגל אליה בעתיד. לאחר מכן עליך לקבוע את ה-SETTING. שם תקבע את בהירות הטקסט, איכותו, כיוון הקריאה (כעמוד רגיל בגודל 8.5x11, ששורותיו נקראות במאונך לחלון המלבני המוארך --PORTRAIT-- או כעמוד ששורותיו מקבילות לצד המוארך של החלון --LANDSCAPE--), וכן את גודל החלון הנסרק.

בשלב השני של פעולת התוכנה הינך עובר לחלק הניתוח --ANALYZE, שבו עליך לקבוע את איתור הטקסט, צורתו ותכונותיו (כגון טורים, טקסט והערות שוליים, וכדומה) כדי לסייע בקריאה נכונה של הטקסט. התוכנה מציעה ניתוח אוטומטי לבחירתך ואזי היא בוחרת את הנתונים בעצמה.

אחת ההתפתחויות המעניינות ביותר--והחשובות ביותר-- בתחום תוכנות המחשב בלשון העברית והלועזית הופיעה לאחרונה בשוק התוכנות הישראלי. השימוש בה עשוי לחולל מהפכה רבתי בנושאי ההקלדה וההוצאה לאור באמצעות מחשב ולספק כלי-עזר לכל מי שמשמש במחשבים בעברית ובלועזית- לצורך כתיבה, עריכה, הדפסה, הוצאה לאור, ופעולות משרד שכיחות. התוכנה החדשה, תוכנת LIOCR, מאפשרת לסרוק (SCAN) ולזהות אותיות עבריות -- ולועזיות -- באופן אופטי והריהי מסבה טקסט עברי ולועזי מודפס לאותיות-מחשב הנקראות בכמה תמלילים עבריים דו- לשוניים ואנגליים מקובלים לצורך עיבוד הטקסטים מחדש, בלי הידרשות להקלדתם מחדש.

תוכנת סריקה וזיהוי אותיות באנגלית ובשפות המשתמשות בטקסט הלאטיני מצויות בשוק, כאומניפייגי ו-וורדסקן, והן מקובלות ושמיות במידה זו או אחרת. ואילו בעברית, התוכנה שלפנינו, שהיא באמת תוכנה רב-לשונית, הינה אחת המעטות המשמשות גם את השפה העברית -- אף כי לעת עתה בלי ניקוד -- ועומדת עתה לרשות הלוקוח העברי בישראל ומחוץ לה.

תוכנת סריקה מעין אלה תסייענה לכל מי שיש תחת ידיו חומר מודפס או חומר שתוקתק במכונת כתיבה בעידן שלפני המחשב, והוא רוצה להמירו לטקסט תמלילני לכל דבר. מעתה הוא יוכל לסרוק את החומר המודפס ולקבלו מחדש בקבצי מחשב לשם עיבודם ותיקונם כטקסט שהוקלד במקורו במחשב באחת התוכנות המקובלות לעיבוד תמלילים בעברית ובאנגלית. אין, איפוא, צורך להקליד את כל החומר מחדש. עליך רק לסרוק את החומר המודפס באמצעות התוכנה, המעבדת אותו לטקסט ממוחשב. בדרך זו תחסוך מאות ואלפי שעות עבודה... אך, כמובן, עליך להשקיע זמן בלימוד התוכנה והשימוש בה. וגם זו תורה ולימוד היא צריכה... ובצד העמל -- גם השכר והתועלת.

בית התוכנה שפיתח את הסורק והמזהה האופטי הוא 'ליגטורה', בירושלים, שמנהלו הוא גדעון בן-צבי, והתוכנה נקראת LIOCR, שפירושו: Optical Character Recognition Ligature. על כך ייאמר: כי מציון תצא תורה... ואף אותה -- את התורה -- אפשר עתה לקרוא, לאחר סריקה, באמצעות התוכנה ב...קריאת אותיות הכתובות בכתב סופר...

לשם הפעלת התוכנה נדרש מחשב מאחד הסוגים הבאים: תואס-י.ב.מ. מסוג AT, מחשב י.ב.מ. PS2, או תואס שבו מותקן

קיימת אפשרות לקבוע את העמוד הנסרק לפי צורתו, אם הוא מעוצב בטורים, או לבחור את הטקסט כטור אחד. אם יש לך כמה מסמכים בעלי צורה קבועה יכול אתה לקבוע מראש את פורמט הקריאה בסטנדרט קבוע, ולשמור את הפורמט הזה לשימוש בעתיד.

התוכנה מספקת כמה כלי-עזר יעילים לטיפול בטקסט הנסרק. יש לך אפשרות לבחור מתוך טור או עמוד חלקי טקסט לשם טיפול וסריקה. יש אפשרות לשנות סידרם של קטעים נסרקים, להשמיט קטעים או לגזור ולהעביר אותם, לשנות טקסט ורקע משחור על לבן ללבן על גבי שחור, וכן להגדיל קטעים.

בשלב השלישי, שהוא שלב זיהוי האותיות, הנקרא OCR (שהוא, כאמור, זיהוי אותיות אופטי), עליך לקבוע את דרך זיהוי האותיות. קיימות שתי דרכי זיהוי. האחת, על-פי מתכונת קבועה מראש הנקראת OMNIFONT שלפיה מזהה התוכנה גופנים (פונטים) של אותיות בלאטינית ובעברית בלא התערבותו של המשתמש. התוכנה מזהה בעצמה את האותיות הנסרקות בכל גדליהן ובכל דרכי הדפסתן לאחר קביעת השפה. יש אפשרות לזיהוי כמה פונטים בעת ובעונה אחת. בבחירת הפונטים ניתן לסרוק גם טקסט של אותיות שנדפסו במדפסת סיכות רגילה.

דרך שנייה מאפשרת למשתמש לאמן את התוכנה לזהות אותיות בעלי סגנון עיצוב מיוחד או שונה. התוכנה מאפשרת פעולה אינטראקטיבית עם מערכת זיהוי האותיות בשלוש דרכים: בדרך האוטומטית, שבה אינך נדרש לפעול כלל, והתוכנה משרבבת סימן מיוחד במקום אות שאינה יכולה לזהותה. בדרך השנייה, ליגטורה נועצת בך כל-אימת שאינה יכולה לזהות אות כלשהי, ועליך לסייע לה בזיהוי האות על-ידי הקשה על לוח המקשים וקביעת האות הנכונה. בדרך השלישית, ליגטורה תעצור בכל אות ותראה לך את זיהויה לשם אישורך. דרך זו נבחרת בדרך כלל לשם אימון התוכנה לזהות פונט מיוחד שאינו מוכר לה.

אימון התוכנה לקריאת חומר לפי 'הזמנה מיוחדת'

לצורך זיהוי טקסט שיש בו אותיות מיוחדות, סימנים מיוחדים, או שפה שאינה נתמכת על-ידי התוכנה, יכול אתה ללמד את ליגטורה לזהות פונטים מיוחדים בשיטה של 'הזמנה מיוחדת', CUSTOMIZED. אתה תלמד את ליגטורה לזהות כל אות ואות, תוכל לעדכן את הזיהוי לפי הצורך, ולשמור את הקובץ לשימוש בעתיד לשם סריקת חומר שנדפס באופן דומה. באימון התוכנה לזיהוי האותיות תוכל לבחור סף מעבר לוודאות הזיהוי. בית התוכנה ממליץ לבחור בתחילה את סף הזיהוי 90 (מתוך 50 עד 100), כדרגת ודאות בזיהוי האות. בהמשך אימון התוכנה, ניתן לשנות את סף הוודאות. התוכנה מאפשרת להפריד בין שתי אותיות דבוקות זו לזו, תופעה הידועה בשם LIGATURE, כשם התוכנה, ולזהות כל אחת מהן בנפרד.

לפני תחילת הסריקה יכול אתה להגביל את השימוש לסריקה בלבד, והעיבוד והעריכה ייעשו בפעם אחרת. הסריקה יכולה להעשות אוטומטית ברצף וללא הספקה, לפי מספר העמודים הנדרש, וכן יכול אתה לקבוע את שמות הקבצים – כל עמוד כקובץ בפני עצמו – שבהם יישמר החומר הנסרק –

שמירה אוטומטית אף היא – כאותיות מחשב מזהות, לשם עיבודו בעתיד. בנוסף על כך עליך לקבוע את מעבד התמלילים שלפי מתכונתו יישמר הקובץ הנסרק וישבו יעבוד אחר-כך. כמה מן התמלילים העומדים לרשותך: איינשטיין, וורד סטאר, א"ב, ועוד... בעברית, וכן רוב התמלילים הלועזיים כוורד פרפקט, וורד, זיירייט, לוטוס וזולתם.

עריכת הטקסט הוא השלב האחרון. העורך של ליגטורה עובר במהירות על פני הטקסט לשם עריכתו ועיבודו, ומציג לפניך אשנב תיקון עם לחיצה כפולה על העכבר, ובאשנב תופיע המלה המקורית – לשם תיקונה. העורך מאפשר כל פעולות העריכה: כהעתקה, העברה, השמטה והדפסת החומר. יש אפשרות לחיפוש והחלפת מלים או סדרת מלים, הדגשת אותיות, וכדומה. כמובן, יכול אתה להעלות את הקובץ במעבד בתמלילים שבחרת ולתקן את הטקסט לפי הצורך. אך השימוש בעורך של ליגטורה יסייע לך לדגל אוטומטית לאותיות מסופקות שהתוכנה לא זיהתה אותן, מה שלא יעשה באופן אוטומטי בתמלילן אחר. מאידך גיסא, העורך של ליגטורה עדיין איננו כולל תוכנת איות והגהה, ככמה מן התוכנות הלועזיות המצויות בשוק.

במקום לעבור על שלושת השלבים הנזכרים בזה אחר זה תוך בחירתם, תוכל ללחוץ על הכפתור 'הרץ' RUN- והתוכנה תעבור על פני השלבים הנדרשים בזה אחר זה. בסריקה יש להיזהר להציב את עמוד הטקסט בדיוק בחלון הסורק, על מנת שהשורות תיקראנה כהלכה. עמוד שהונח עקום לא ייקרא כהלכה.

בדקתי תוכנה זו במשך זמן ממושך, שבו עיבד בית התוכנה כמה וכמה גירסאות, ולאחרונה הגירסה 2.1 לחלוטות. סרקתי טקסטים באיכויות הדפסה שונות כדי לבחון את יעילות התוכנה. המסקנה הכללית המתבקשת מכל נסיונותיי היא, שהתוכנה יעילה ביותר כאשר ניתן לה לסרוק טקסט בעל איכות הדפסה מעולה, אותיות ברורות ונקיות, מודפסות בצבע דפוס שחור על גבי נייר לבן. נייר עיתון, שאינו לבן, עלול לשבש את איכות הסריקה. התוכנה מצטיינת ביכולת ובגמישות לזהות מגוון רחב של אותיות מעוצבות בסגנון שונה בלא הצורך ב'אימון' מיוחד. ה־omnifont recognition שלה הוא מעולה. התוכנה אמורה לשמור את הדגשות הטקסט המקורי: אותיות נוטות, שמנות ומקוקקות.

בסריקת טקסט של מכתב בן עמוד, שנדפס במקורו באות נרקיס במדפסת לייזר, נשתבשו האותיות טיוס' זו בזו, הספרה 1 זוהתה כאותן, והאנגלית לא נקראה כהלכה. התוצאה היתה: שלוש שגיאות בעמוד, שהיא תוצאה די טובה. רוב השגיאות נובעות מדמיון האותיות: האות כ' נקראה משום מה כ-ס' עם סף ודאות של 90%. ו' נוטה להיות מזוהה כ'ן, מ' כ-ס', ה' כ-ח'. בסריקת טקסט מתוך ספר מודפס שיבש הסורק אותיות ח' כ-ה', ל' כ-ע' (לומר' נקרא 'עומר', ילא' – 'עא'). בעמוד זה היו 15 שגיאות, ברובן החלפת ל' ב-ע', ו-ה' ב-ח'. עמוד אחר נסרק מתוך כתב-עת מודפס ובו נשתבשו כארבע מלים, ואף זו תוצאה מצויינת, לדעתי. בעמוד אחר, מתוך מאמר מדעי, נפלו שיבושים בזיהוי האנגלית שהיתה משובבת במסגרת העברית. רוב השיבושים נפלו בהערות השוליים, שנדפסו במקור באות קטנה יותר, ומשום כך קשה יותר לזיהוי. האות ח' נקראה כ-מ' בטקסט עצמו ורוב השגיאות נבעו מקריאה משובשת זו.

מתוך שלושים שורות של מאמר שנדפס תחילה ב"הדואר"

אשר נסרק בטורו הראשון, נפלו כ-17 שגיאות, חלקן בשיבוש כי ר"ב. הטקסט המקורי לא היה מודפס באופן ברור והאותיות לא היו גדולות ובהירות, ומשום כך מספר שגיאות רב יותר. עם זאת, הטקסט מוקלד עתה בהקלדת מחשב, ולאחר תיקון מהיר של 17 השגיאות, הריהו מוכן לתיקון, עיבוד והדפסה מחדש, בלא העסקתה של קלדנית שבוודאי היתה מכניסה יותר מ-17 שגיאות...

מהי יעילות התוכנה? זו היא השאלה הרלוונטית ביותר. דהיינו, האם משתלם בכלל להשתמש בתוכנה זו, במקום למסור את החומר להקלדה מחדש? כל אחד יחליט בעצמו, לפי צרכיו. אך ברור, שככל שהטקסט המקורי לסריקה מודפס היטב וברור, כן תמעטנה השגיאות. בהשוואה לתוכנות לועזיות מקובלות, ליגטורה איננה נופלת מן הטובות שבהן בתחום.

במגוון הביצועים האפשריים, בגמישות מרחב הזיהוי וביעילות. אך אם בידך כתב-יד, מומלץ למסור אותו לקלדנית, ניסיתי לסרוק כתב-יד מן המאה הי"ח כדי לראות אם אוכל לקבל טקסט מוקלד, ומובן, שהתאכזבתי... אך נקווה, שלא ירחק היום ובית התוכנה יאפשר לזהות גם כתב-יד.

אפשר לפנות ישירות אל בית התוכנה בירושלים בפקסימיליה מספר 513395-2-972, או בכתב: ליגטורה, מפעלי בן-צבי, רח' בית הדפוס 11, איזור התעשייה בגבעת שאול, ירושלים 95483. אפשר גם לפנות אל משרדה החדש של ליגטורה בבוסטון אל Ed Mc Guiggan, טלפון 617-238-6734. מספר הפאקס: 617-272-3085. מחיר התוכנה עם אפשרות לקריאה בעברית \$1,195.

אוניברסיטת מרכז פלורידה, אורלנדו, פלורידה