מדריך למשתמש

HOCR

2011

Maxim Drabkin

Niv Maman

*אוניברסיטת חיפה*

*החוג למדעי המחשב*

# תוכן העניינים

[הסבר כללי](#_הסבר_כללי) .................................................................................................................. 2

[התקנת קבצים](#_התקנת_הקבצים).............................................................................................................. 3

[זיהוי טקסט](#_זיהוי_טקסט) ................................................................................................................. 4

[זיהוי תמונה](#_זיהוי_תמונה) ..................................................................................................... 4

[זיהוי PDF](#_זיהוי_PDF) ........................................................................................................ 5

[עיבוד תמונה](#_עיבוד_תמונה) ............................................................................................................... 5

[סיבוב](#_סיבוב) ............................................................................................................. 5

[ניקוי](#_Clean) ............................................................................................................... 6

[חלוקת שחור-לבן](#_Threshold) .............................................................................................. 6

[מרכוז](#_Centerize) ............................................................................................................. 6

[שימוש נפוץ](#_שימוש_מומלץ_בפילטרים)...................................................................................................... 6

[אימון רשת הניורונים](#_אימון_רשת_הניורונים)...................................................................................................... 7

[יצירת פונטים מתוך תמונה](#_יצירת_פונטים_מתוך).................................................................................. 8

[יצירת פונטים רקים](#_יצירת_פונטים_רקים)............................................................................................ 8

[אימון רשת שנבחרה](#_אימון_רשת_שנבחרה).......................................................................................... 8

[תיקון שגיאות](#_תיקון_שגיאות).................................................................................................... 9

[הרצת Script ליצירת ואימון הפונטים](#_הרצת_script_ליצירת)................................................................................. 9

[Create All Fonts](#_Create_All_Fonts)............................................................................................... 9

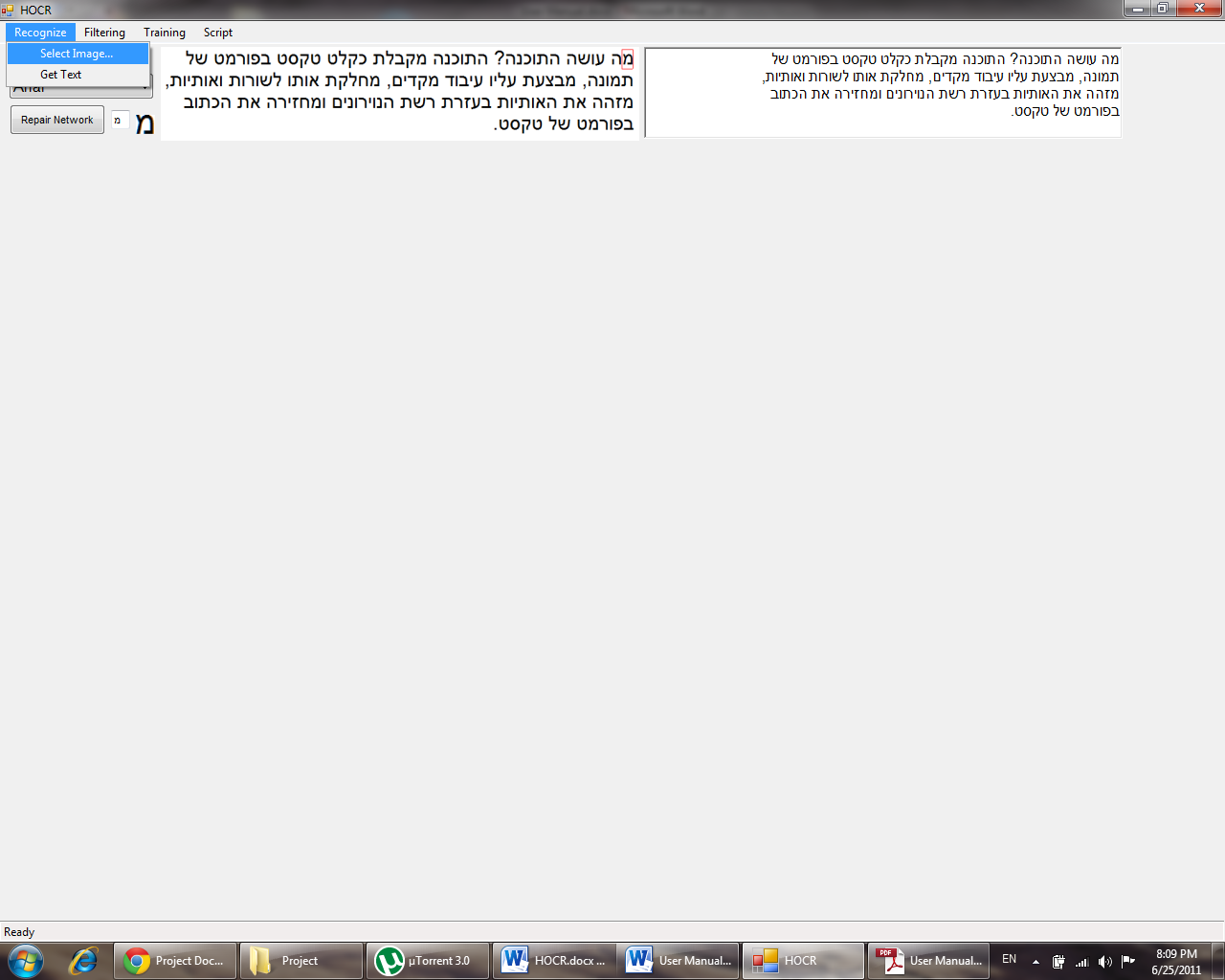
[Train All Fonts](#_Train_All_Fonts)................................................................................................. 9

# הסבר כללי

המערכת מורכבת משלושה חלקים עיקריים:

### זיהוי – Recognize

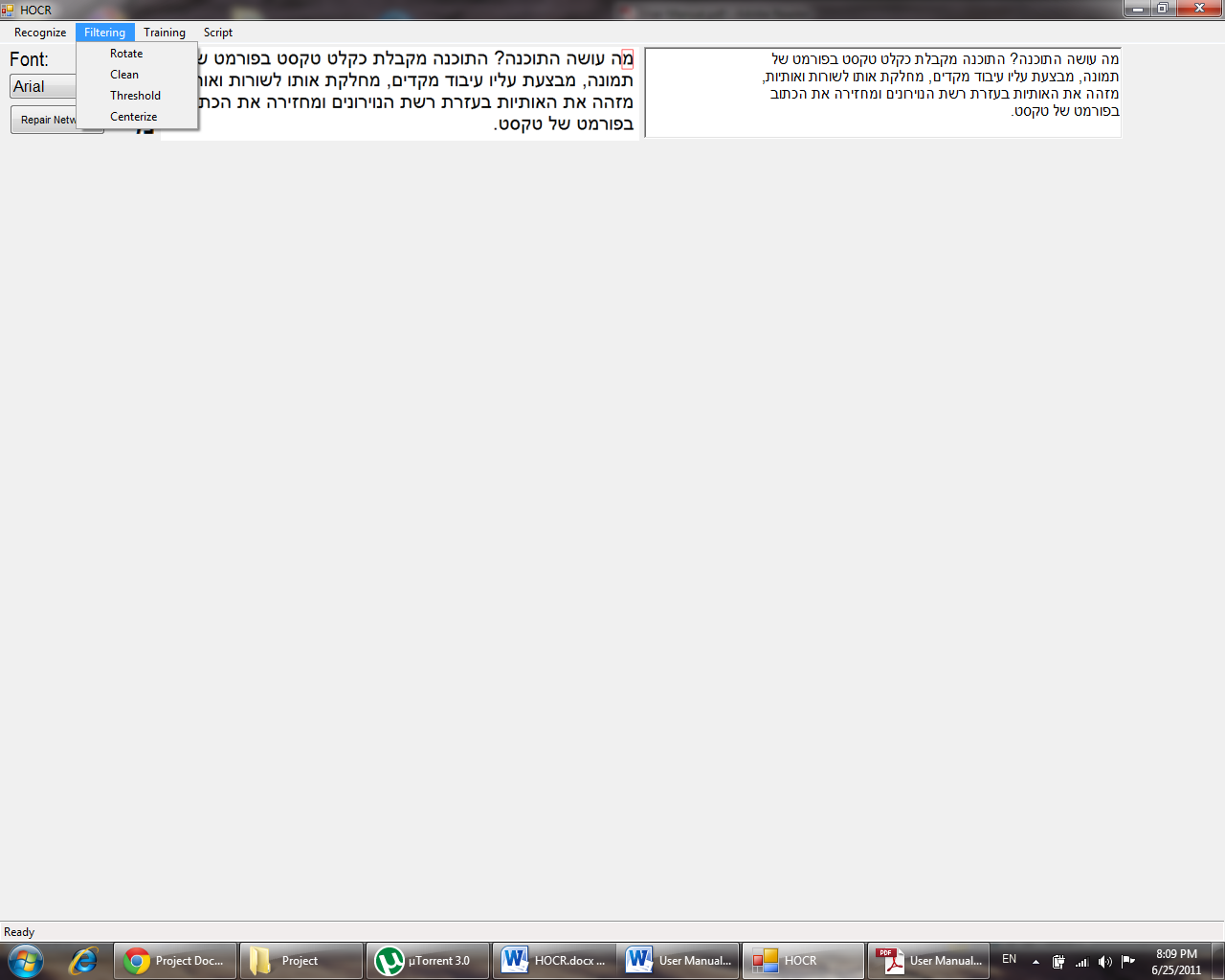
הקומפוננטה האחראית על בחירת תמונה או קובץ PDF, וכמו כן על קבלת הטקסט מהתמונה שנבחרה.



### עיבוד תמונה – Filtering

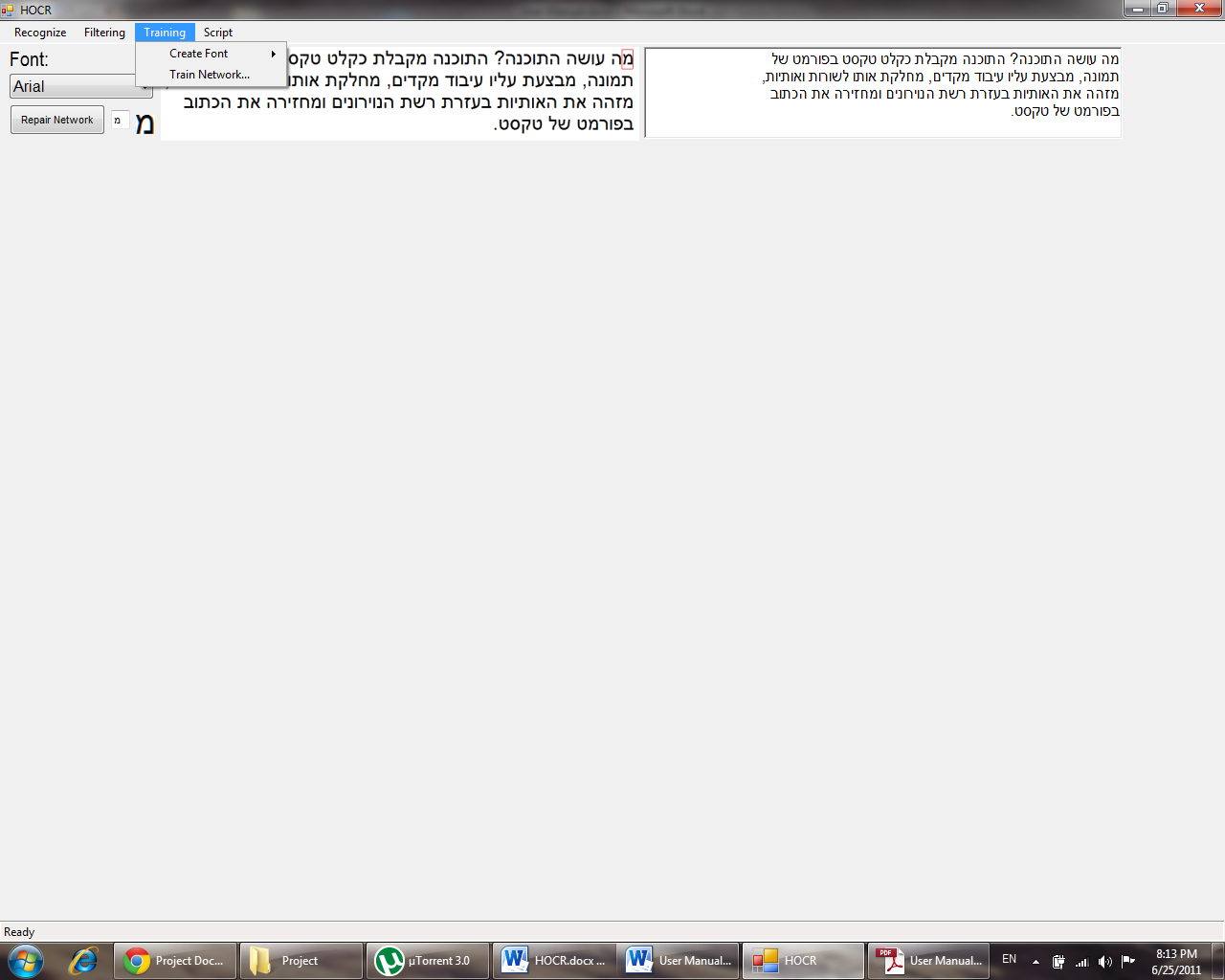
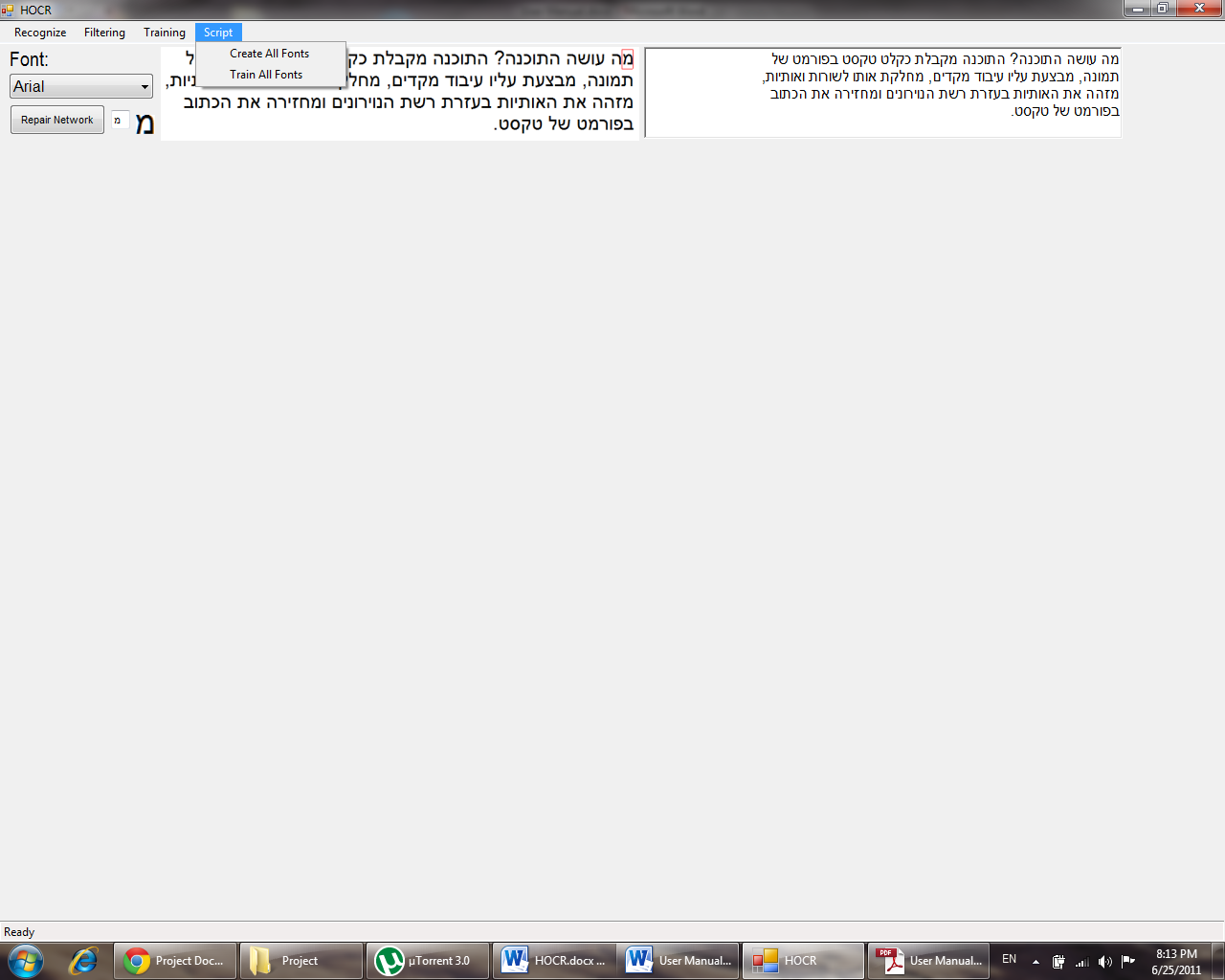
הקומפוננטה אשר אחראית על ביצוע עיבוד לתמונה שנבחרה.

מאפשר ליישר את התמונה אם הטקסט בה אלכסוני, לנקות אותה מרעשים, להוריד קצוות לבנים על מנת להתרכז רק בטקסט.



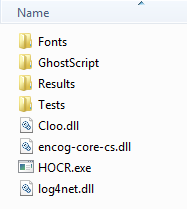
### אימון רשת ניורונים - Training & Script

הקומפוננטה האחראית על הרשת הניורונים ועל הלמידה בעזרת Back Propagation.

# התקנת הקבצים

על מנת להריץ את המערכת, יש ליצור תיקייה במחשב הלוקלי ולהעתיק אליה את הקבצים הבאים:



DLL that included in the neural network package

The main executable of our application

DLL responsible on neural network classes

DLL that included in the neural network package

Folder that holds several examples

Folder where all the results will be saved

GhostScript folder to handle PDF reading

Folder that holds all the learned fonts

על מנת להריץ את האפליקציה, יש לפתוח את הקובץ HOCR.exe

## תנאים מוקדמים להרצה

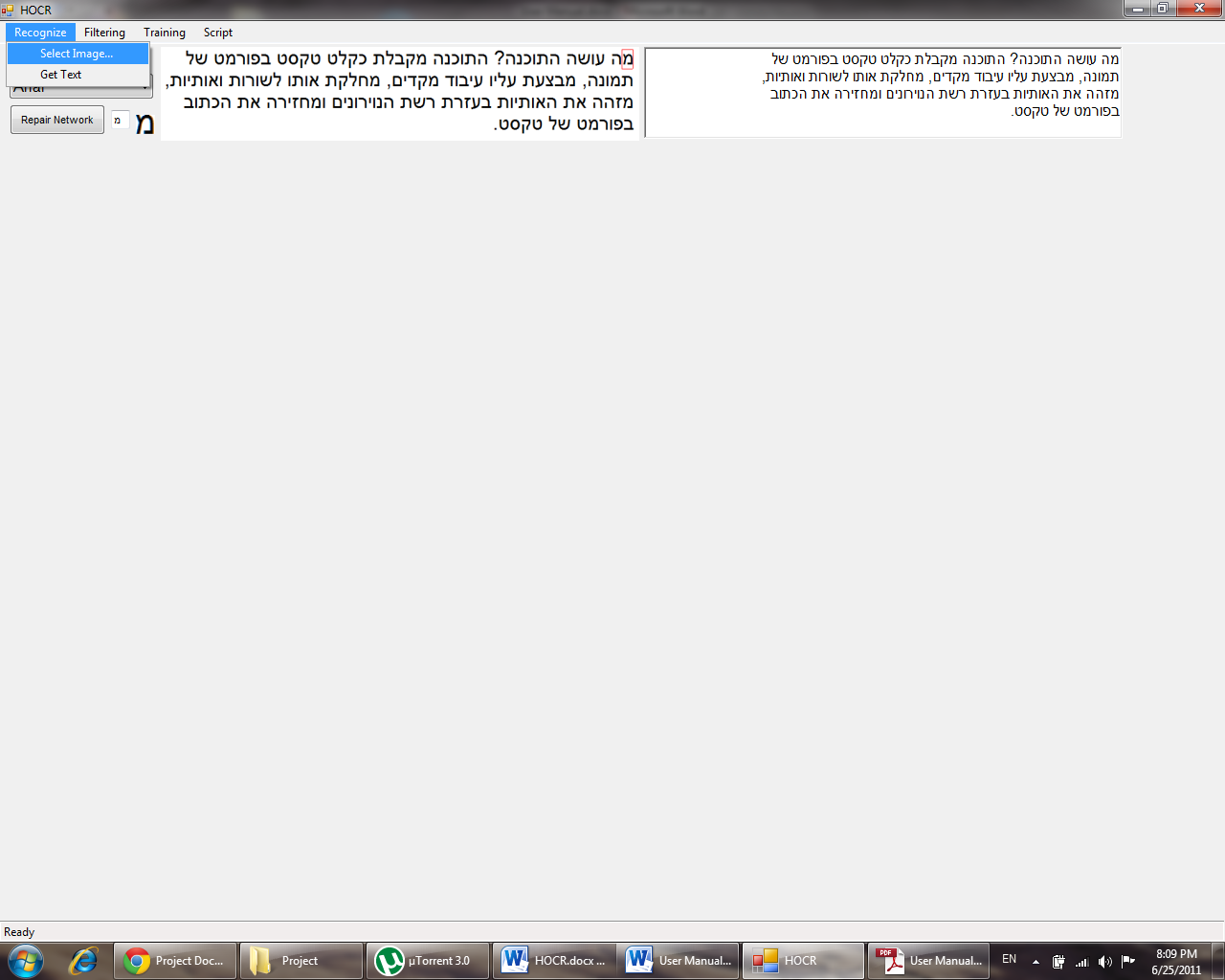
על מנת שהאפליקציה תעבוד, יש לוודא שעל המכונה מותקן:

* אם סביבת העבודה היא Windows, יש להתקין .NET Framework
* אם סביבת העבודה היא Linux, יש לוודא שמותקן MONO.

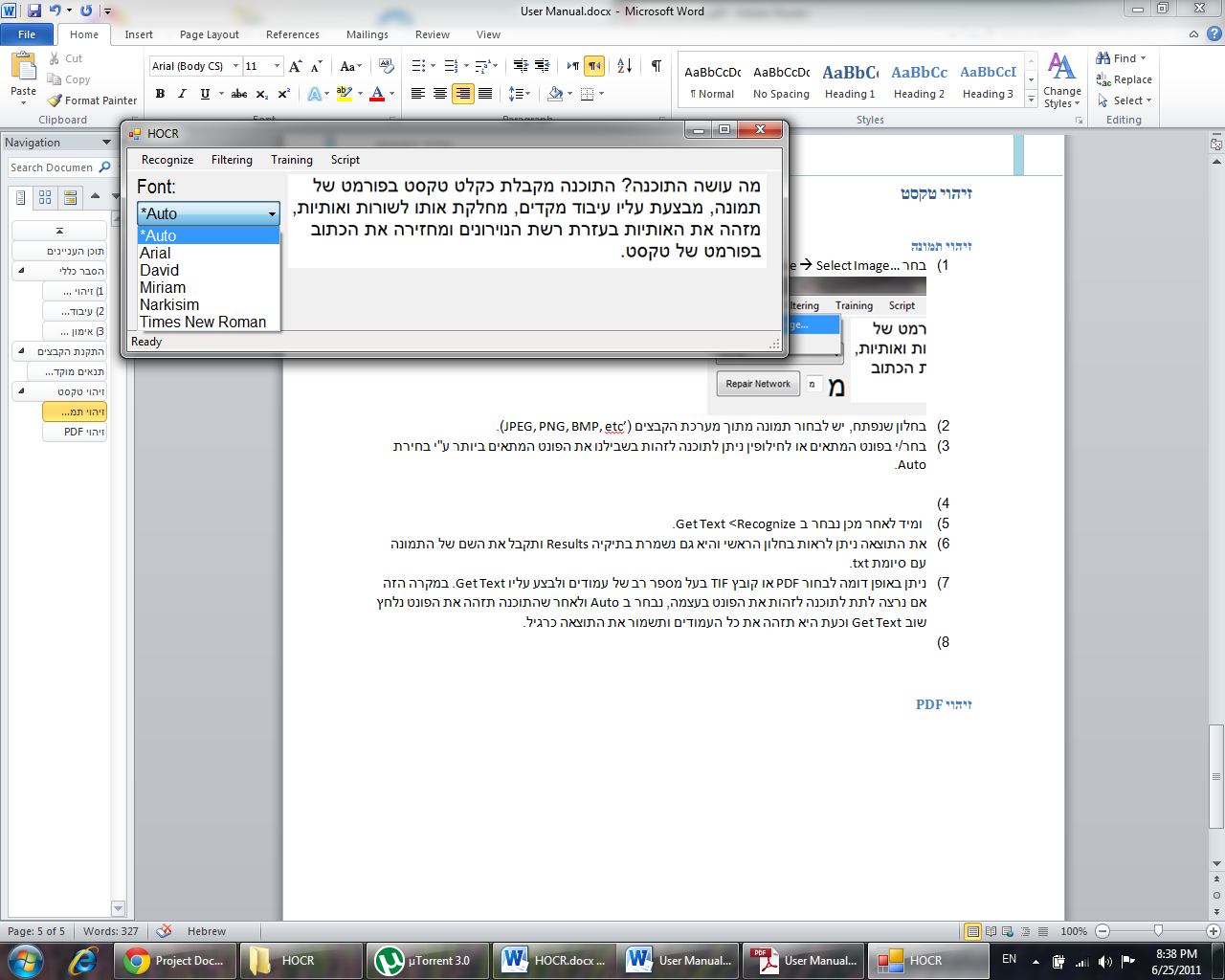
# זיהוי טקסט

### זיהוי תמונה

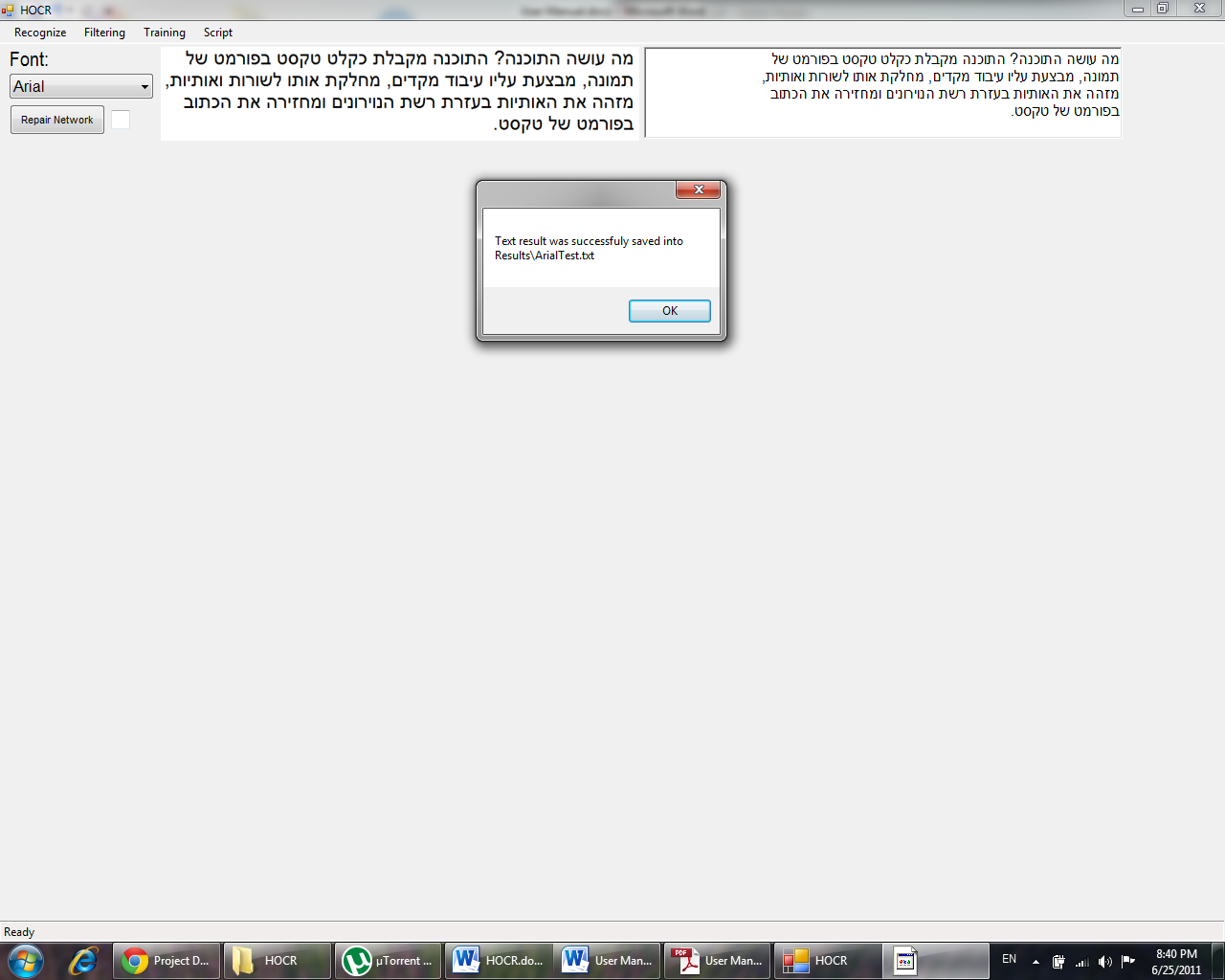
1. בחר/י באפשרות Recognize 🡪 Select Image…



1. בחלון שנפתח, יש לבחור תמונה מתוך מערכת הקבצים (JPEG, PNG, BMP, etc’).
2. בחר/י בפונט המתאים או לחילופין ניתן לתוכנה לזהות בשבילנו את הפונט המתאים ביותר ע"י בחירת Auto.



1. בחר/י באפשרות Recognize 🡪 Get Text
2. את התוצאה ניתן לראות בחלון הראשי והיא גם נשמרת בתיקיה Results ותקבל את השם של התמונה עם סיומת txt.



### זיהוי PDF

ניתן באופן דומה לבחור PDF או קובץ TIF בעל מספר רב של עמודים ולבצע עליו Get Text. במקרה הזה אם נרצה לתת לתוכנה לזהות את הפונט בעצמה, נבחר ב Auto ולאחר שהתוכנה תזהה את הפונט נלחץ שוב Get Text וכעת היא תזהה את כל העמודים ותשמור את התוצאה כמו בסעיף הקודם.

# עיבוד תמונה

במידה ולתמונה פגמים ויזואלים (לא ישרה או רועשת מדי) יש להשתמש בפילטר המתאים.

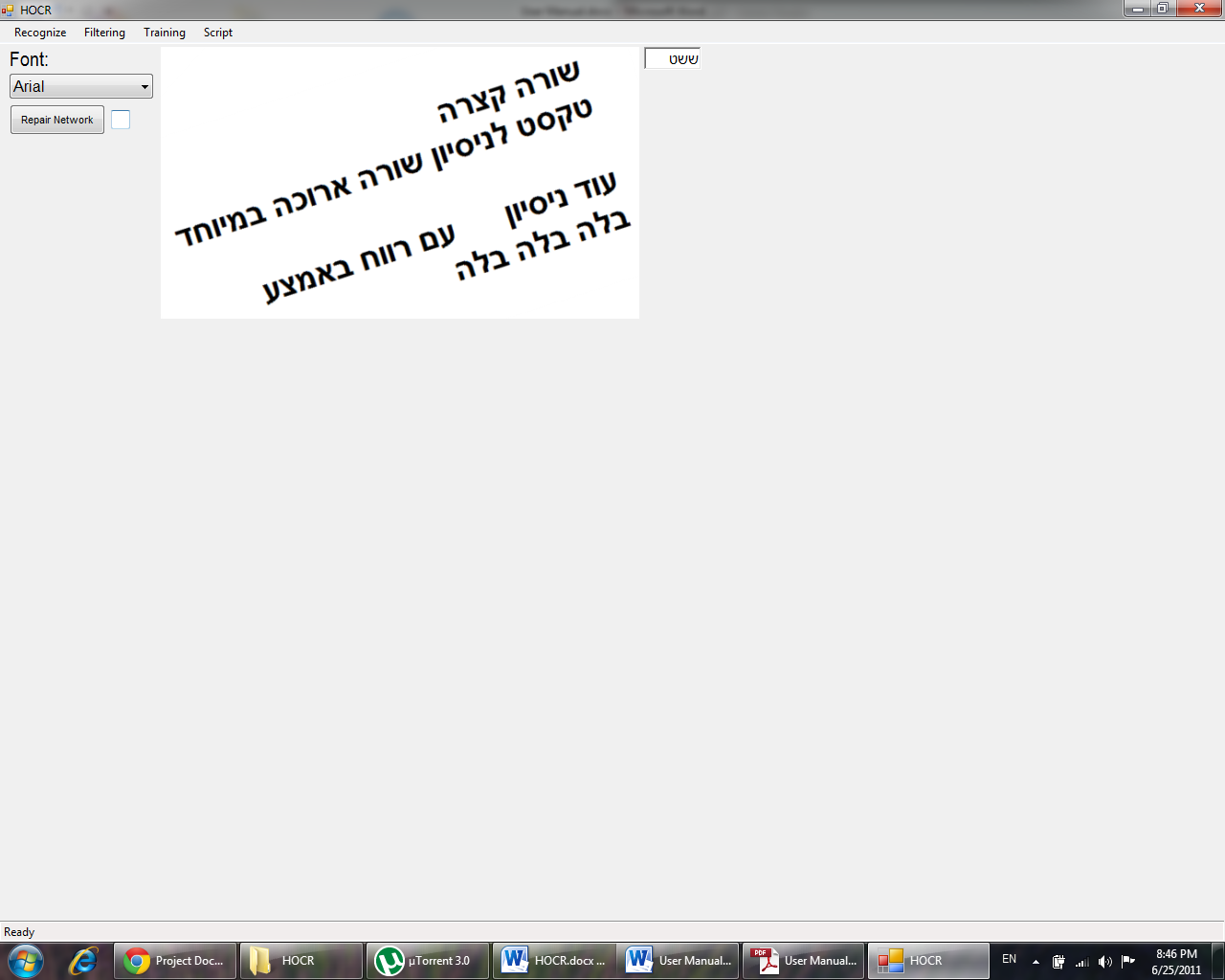
ישנם 4 פילטרים מובנים בתוך התוכנה: Rotate, Clean, Threshold, Centerize והמשתמש יכול להשתמש בהם במידת הצורך.

### סיבוב

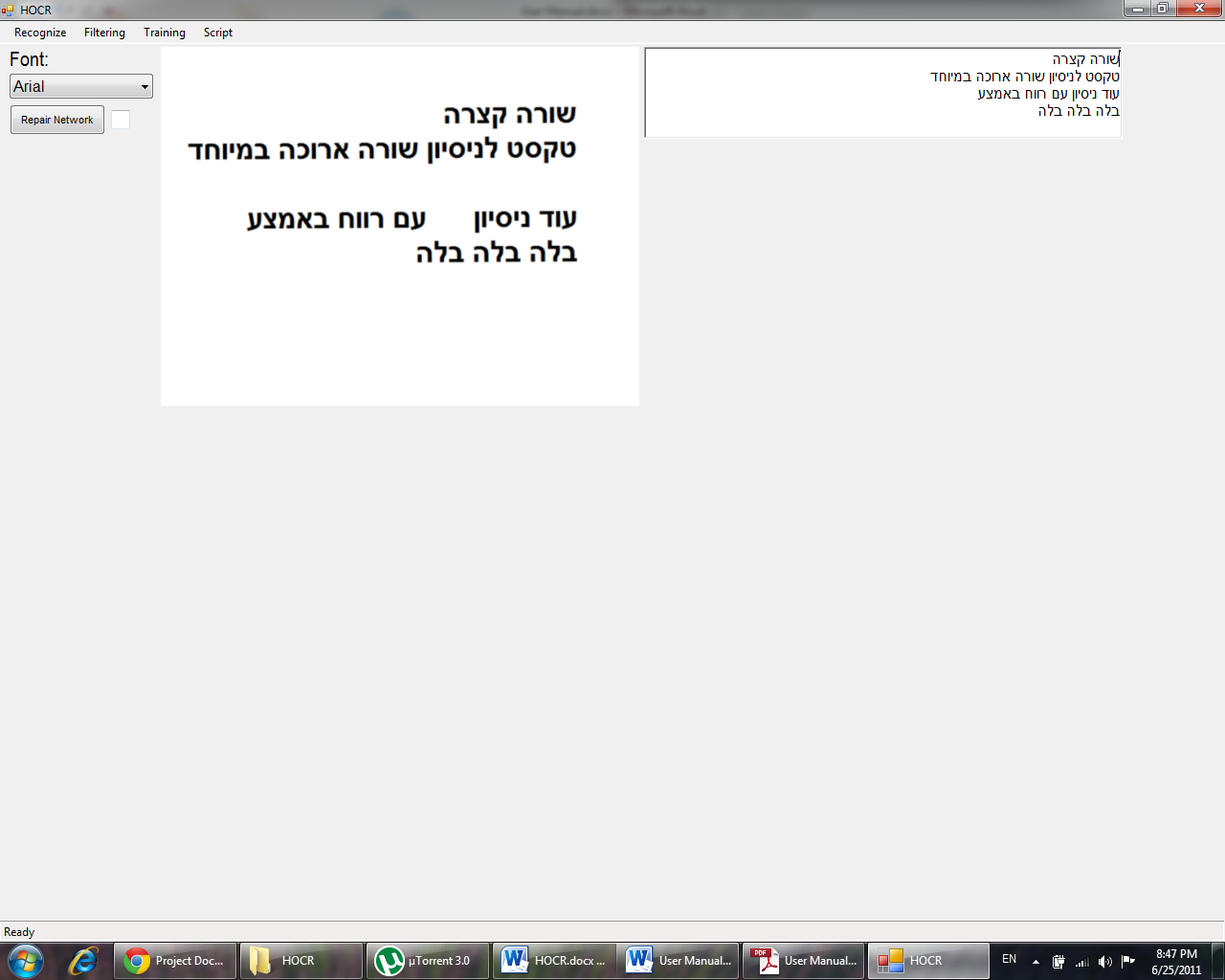
בהנתן תמונה שנסרקה באלכסון, בחירה בפילטר זה תפעיל אלגוריתם ליישור הטקסט, כך שבסופו של דבר הזיהוי התמונה שנצטרך לזהות תהיה ישרה.

על מנת להפעיל את הפילטר, יש לגשת ל- Filters🡪 Rotate

דוגמא לזיהוי שגוי כתוצאה מתמונה מסובבת:



לאחר הפעל הפילטר נקבל תמונה ישרה, וכתוצאה מכך גם זיהוי מוצלח:



### Clean

בהנתן תמונה שנסרקה מעיתון או פשוט מכילה הרבה רעש, ניתן לנקות את הרעש בשביל לאפשר זיהוי אופטימלי של התווים.

על מנת להפעיל פילטר זה, יש לבחור באפשרות Filtering 🡪 Clean

### Threshold

לכל תמונה סרוקה, גם אם היא נראית כתמונת שחור-לבן, ישנם גוונים שונים של אפור, בעיקר בקצוות של האותיות.

על מנת לאפשר זיהוי אופטימלי, הלימוד של הרשת נעשה על פיקסים שחורים ולבנים בלבד.

לכן לפני זיהוי, ניתן להריץ את הפילטר הנ"ל על מנת להגדיל את הסיכויים לזיהוי אופטימלי.

על מנת להריץ פילטר זה, יש לבחור באפשרות Filtering 🡪 Threshold

### Centerize

פילטר זה מאפשר להוריד את השוליים הלבנים מהתמונה וממרכז את הפיקסלים שמכילים את הטקסט.

פעולה זו נועדה כדי לאפשר זיהוי מהיר יותר, שכן אין צורך לעבור על פיקסלים מיותרים.

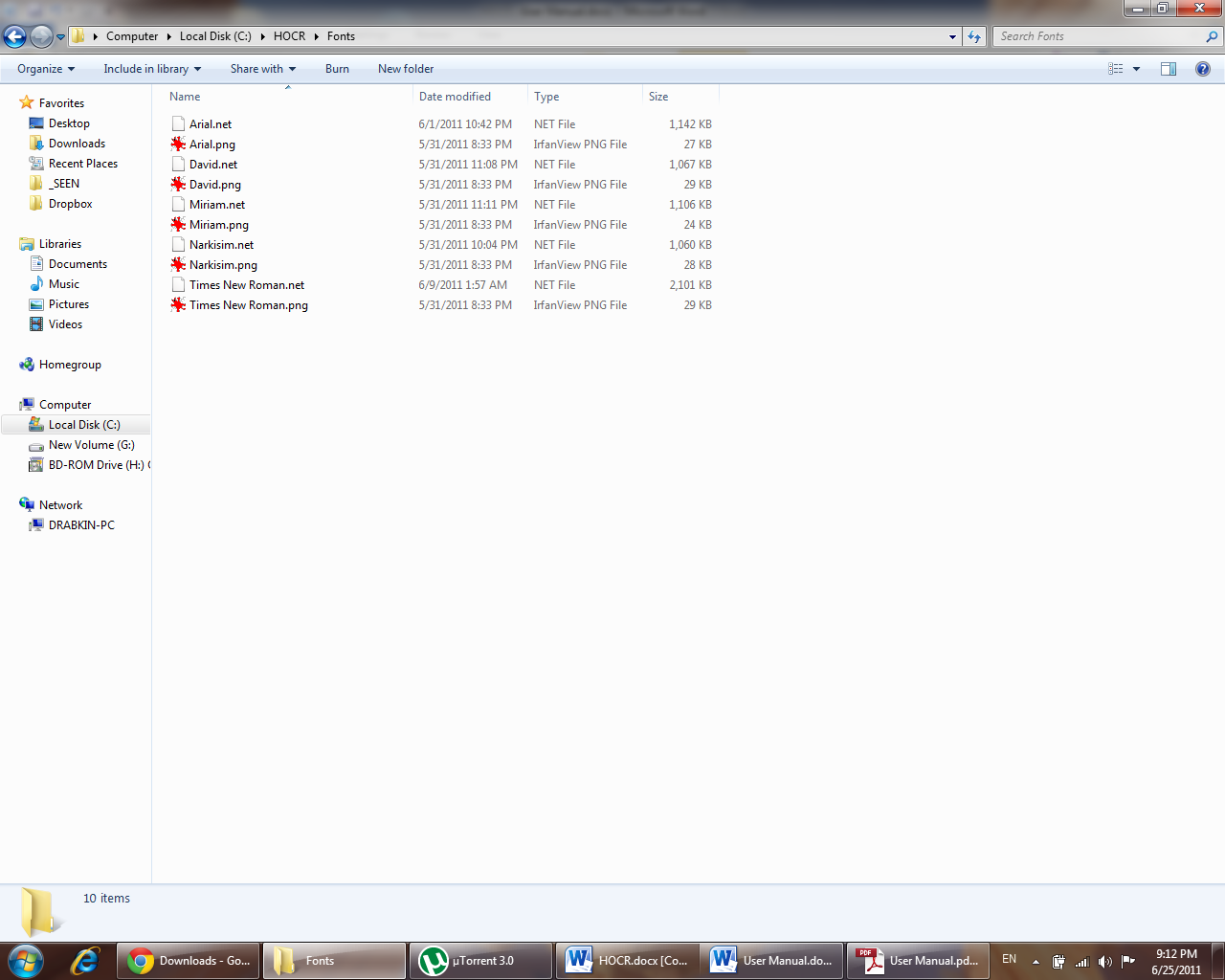
על מנת להריץ פילטר זה, יש לבחור באפשרות Filtering 🡪 Centerize

## שימוש מומלץ בפילטרים

ברוב התמונות נשתמש ב Rotate 🡨 Threshold 🡨 Centerize בסדר הזה.

# אימון רשת הניורונים

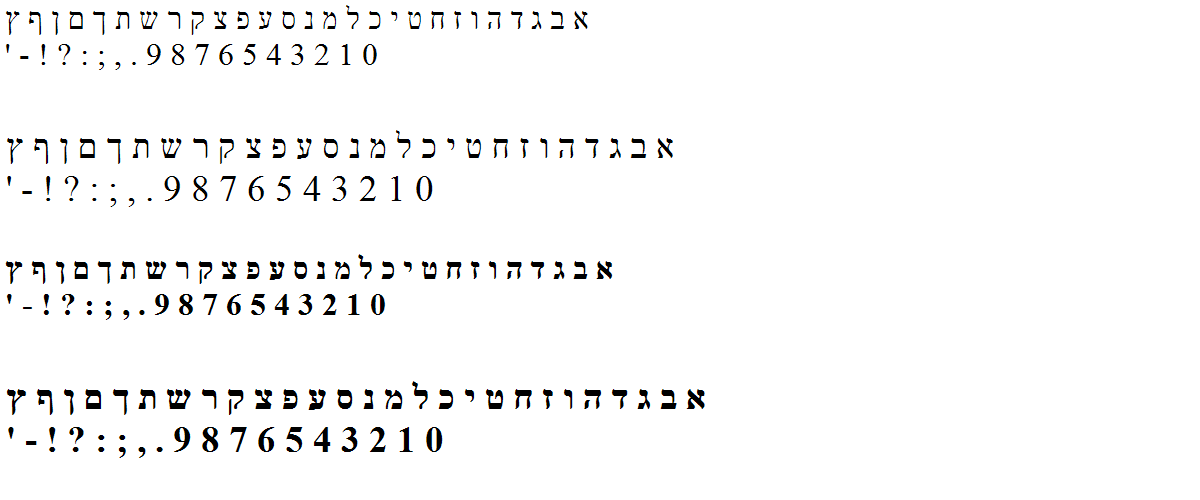
למערכת מצורפים חמשת הפונטים הנפוצים ביותר:



ניתן למצוא אותם בתיקיית Fonts שבתיקייה ההתקנה.

כל פונט מיוצג על ידי 2 קבצים:

1. קובץ PNG- מכיל כמה סדרות של אותיות בעברית, מספרים וסימני פיסוק.



זהו הקובץ שמשמש ללימוד הפונט, כאשר אנו יודעים מהו הסדר של האותיות.

1. קובץ NET- זהו הקובץ אשר מייצג את הרשת הניורונית של הפונט הנוכחי.

קובץ זה הוא reusable, שכן כל תיקון לפונט דרך התוכנה מעדכן את הקובץ, ללא צורך ללמד הכל מהתחלה.

התוכנה בנויה כך שניתן להוסיף כל פונט ובעזרת התוכנה לאמן אותו עד לקבלת תוצאות מספקות.

## יצירת פונטים מתוך תמונה

השתמש/י באופציה זו כאשר יש לך קובץ PNG (כמו שתואר בסעיף הקודם) אשר מכיל את רשימת כל התווים באלף-בית העברי בפונט החדש.

1. בחרי/י באופציה Training 🡪 Create Font🡪 From…
2. בחר/י את התמונה המכילה את כל האותיות בפונט המבוקש.
3. בצע/י בצעדים שמתוארים תחת סעיף [אימון רשת שנבחרה](#_אימון_רשת_שנבחרה)

לאחר סיום הלמידה, בחירה בפונט המבוקש ובתמונה מסויימת המכילה טקסט בפונט הזה, תחזיר את הטקסט המבוקש.

## יצירת פונטים רקים

השתמש/י באופציה זו כאשר תרצה/י ללמד פונט חדש לגמרי, לדוגמא כתב יד.

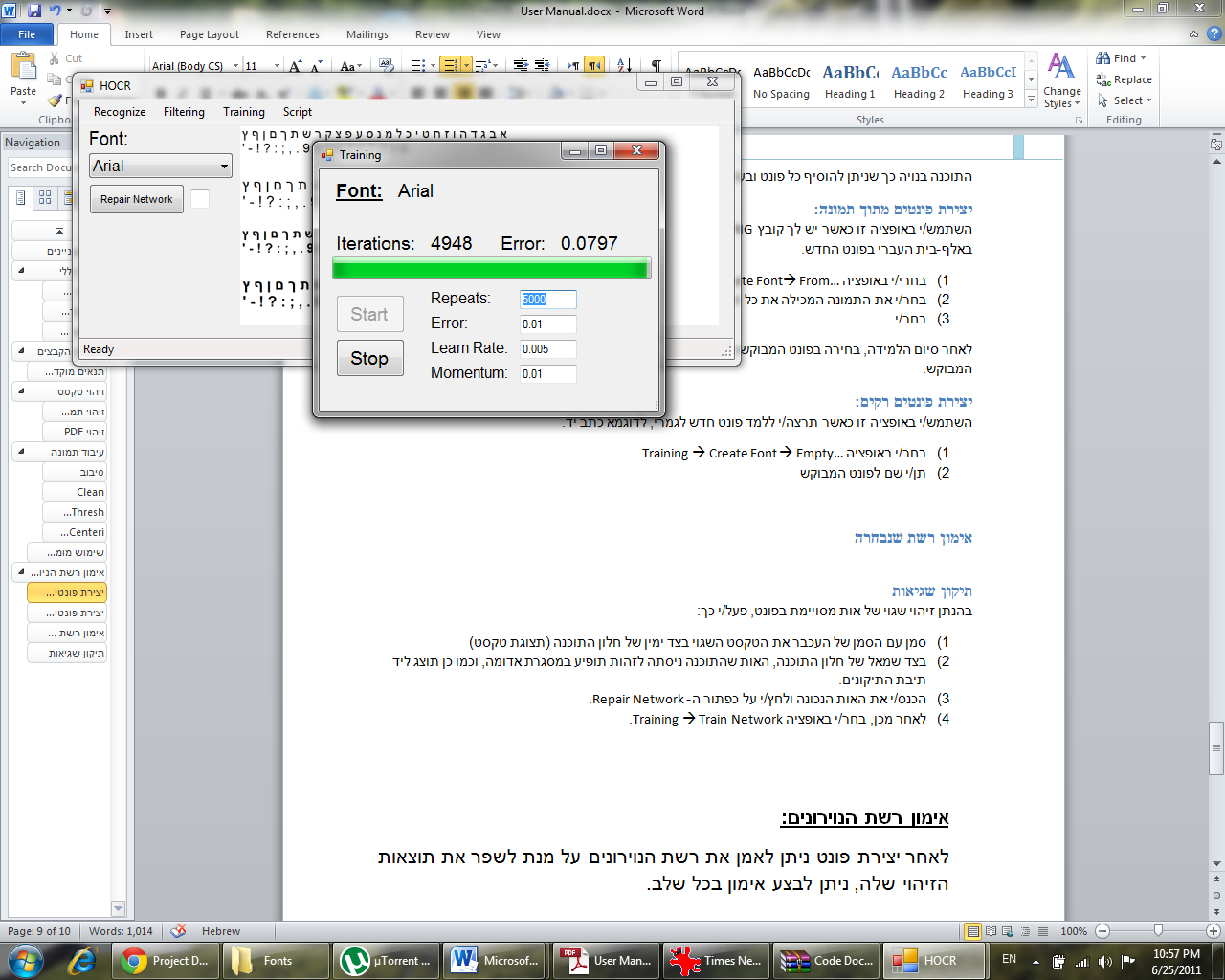
1. בחר/י באופציה Training 🡪 Create Font 🡪 Empty…
2. תן/י שם לפונט המבוקש

## אימון רשת שנבחרה

לאחר יצירת פונט ניתן לאמן את רשת הנוירונים על מנת לשפר את תוצאות הזיהוי שלה. ניתן לבצע אימון בכל שלב.

1. בחר/י בפונט הרצוי בצד שמאל של התוכנה.
2. בחר/י באפשרות Training 🡪 Train Network….
3. בחר/י פרמטרים מתאימים (ניתן להשאיר את ברירת המחדל)
4. לחץ/י START.

כעת ניתן לראות כמה סבבים של אימון ביצענו ומהו אחוז השגיאה הנוכחי ולעצור מתי שנרצה או לחכות שהאימון יסתיים.



כזכור, התוצאה נשמרת בקובץ FONT\_NAME.NET.

**הערה:** קובץ זה ניתן להעביר ממערכת למערכת, ואין צורך לבצע בכל פעם אימון מחדש.

## תיקון שגיאות

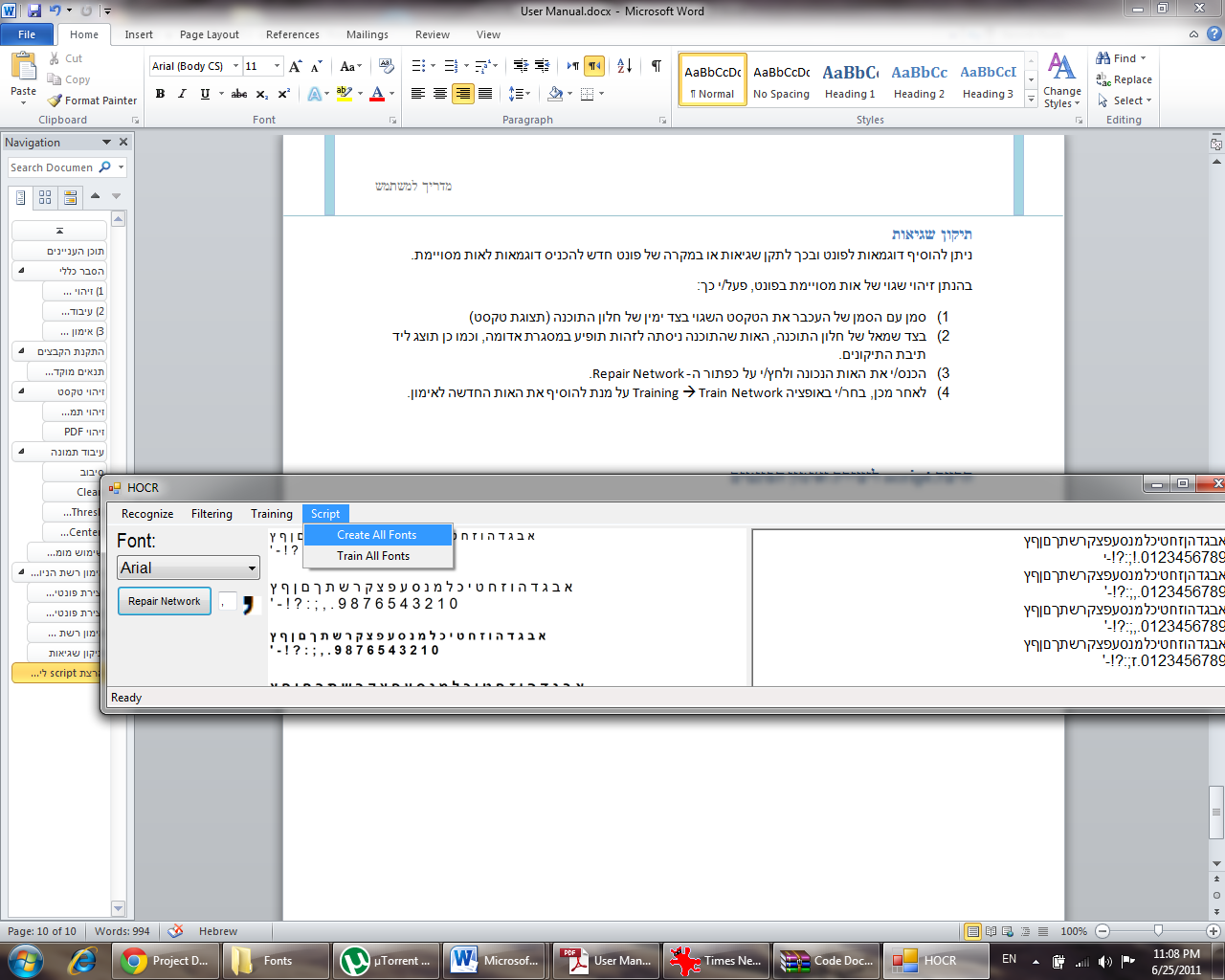
ניתן להוסיף דוגמאות לפונט ובכך לתקן שגיאות או במקרה של פונט חדש להכניס דוגמאות לאות מסויימת.

בהנתן זיהוי שגוי של אות מסויימת בפונט, פעל/י כך:

1. סמן עם הסמן של העכבר את הטקסט השגוי בצד ימין של חלון התוכנה (תצוגת טקסט)
2. בצד שמאל של חלון התוכנה, האות שהתוכנה ניסתה לזהות תופיע במסגרת אדומה, וכמו כן תוצג ליד תיבת התיקונים.
3. הכנס/י את האות הנכונה ולחץ/י על כפתור ה- Repair Network.
4. לאחר מכן, בחר/י באופציה Training 🡪 Train Network על מנת להוסיף את האות החדשה לאימון.

# הרצת script ליצירת ואימון הפונטים

המערכת מכילה שני סקריפטים אוטומטיים שאמורים להקל על המשתמש:



### Create All Fonts

סקריפט זה מייצר את חמשת הפונטים הנפוצים ביותר.

בחירה באפשרות Script 🡪 Create All Fonts תיצור בתיקיית ה- Fonts את חמשת הפונטים (קובץ PNG וקובץ NET).

הערה: מיד לאחר יצירת הפונטים, הם עדיין לא מאומנים והזיהוי ייכשל.

### Train All Fonts

סקריפט זה מאמן את כל הפונטים שמוגדרים במערכת (חמשת הפונטים המובנים וכל הפונטים החדשים שהמשתמש ייצר בעצמו).