פרק תשיעי

כִּימוּת המפגשים (תכונה א)

כאן נעסוק במפגשים של צפני *ELS* על פני טבלאות דו-ממדיות. אנו עוקבים אחרי תכונה א (הנזכרת לעיל):

נטייה למפגש בין צפני *ELS* המייצגים "ביטוי א", לבין צפני *ELS* המייצגים "ביטוי ב", כאשר קיים קשר מושגי מובהק בין "ביטוי א" ל"ביטוי ב".

אנו זקוקים לכלי מדידה כדי לקבוע אם קיימת תכונה כזאת. נתקדם כדלקמן.
נניח שעומד לרשותנו מחשב ובו טקסט נתון ותוכנה מתאימה. עתה יש להקליד למחשב צמד ביטויים – "ביטוי א" ו"ביטוי ב" - שיש קשר מושגי ביניהם.

- ראשית, המחשב סורק את הטקסט, מוצא את המד"שים של "ביטוי א" ושל "ביטוי ב" השייכים ל"אוסף" (כלומר, הם צפני *ELS*).
- אחר כך, המחשב מכין טבלאות דו-ממדיות עבור כל מד"ש שמצא. אם המד"ש מופיע בדילוג d, אזי בונה המחשב מן הטקסט סדרה של טבלאות הנקבעות על פי גודל דילוגו של מד"ש זה: טבלה ובה d טורים, טבלה ובה $[d/2]$ טורים, טבלה ובה $[d/3]$ טורים, וכן הלאה.

בטבלאות אלו יופיע המד"ש בצורה מכונסת. כדי שמפגש בינו לבין המד"ש של הביטוי השני יהיה 'מוצלח', צריך גם המד"ש השני להופיע בצורה מכונסת על פני הטבלה (זו הסיבה שאנו מתעניינים דווקא בטבלאות הנקבעות על ידי המד"שים: שהרי אם קיים מפגש מכונס ביניהם – הוא חייב להתרחש על אחת מן הטבלאות הללו).

בצעד הבא צריך לכַמֵּת (מלשון כמות) את המפגשים בין המד"שים על פני הטבלאות דו-ממדיות שנתקבלו בצעדים הקודמים. דבר זה נעשה על ידי הגדרה של "מידת הקִרבה" בין שני מד"שים. "מידת הקִרבה" נותנת ערך מספרי (כמותי) לכל מפגש בין צמד מד"שים המייצגים צמד ביטויים. אנו מבטאים בכימות זה את המאפיינים הבאים:

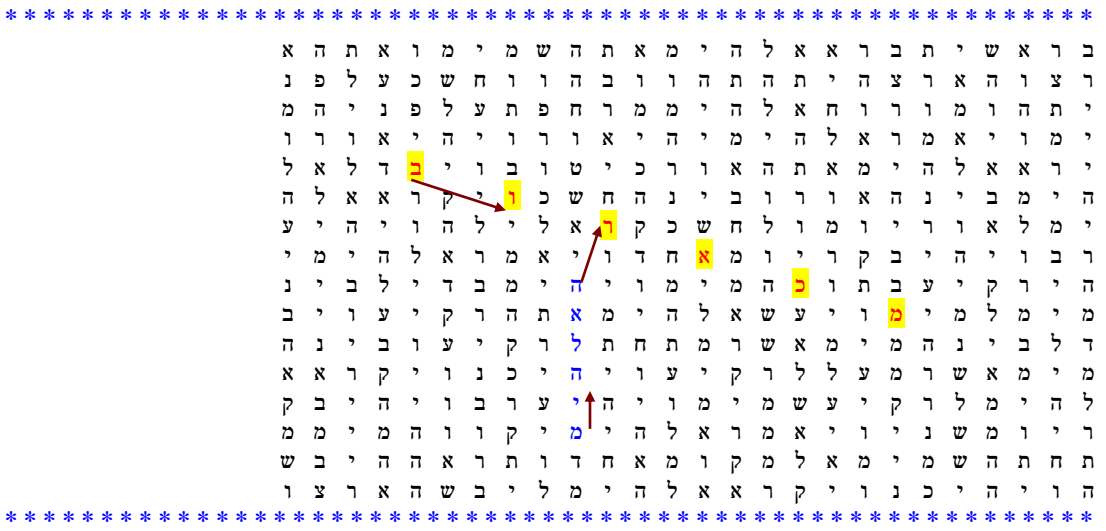
א. טיב הנפגשים: המפגש מתקיים בין צפני *ELS*, דהיינו, מד"שים שהם מינימליים בקטעים גדולים בספר.
ב. טיב המפגש: המפגש נראה כמִקבץ מכונס; כלומר, המד"שים מופיעים על פני הטבלה כאשר הם
   1. קרובים זה לזה.
   2. אינם מפוזרים.

כימות נכון של התכונה שבמעקב צריך לתאר את המרכיבים האלה ולהכיל שני חלקים: חלק המתאר את טיב הנפגשים, וחלק המתאר את טיב המפגש.

חלק הכימות, המתאר את טיב הנפגשים, יעניק ציון "טוב יותר" למפגשים בין מד"שים שהם מינימליים בקטעים גדולים יותר בספר. כך, למשל, הוא יעניק ציון "טוב יותר" למפגש בין צמד מד"שים, שכל אחד מהם מינימלי בכל הספר, לעומת מפגש בין צמד מד"שים שאחד מהם (או שניהם) מינימלי רק במחצית הספר.

חלק הכימות, המתאר את טיב המפגש, יעניק ציון "טוב יותר" למפגשים, שבהם מופיעים המד"שים כשהם קרובים זה לזה ולא מפוזרים. זה החלק המודד את "דחיסות" המפגש. המודד "לדחיסות" לוקח בחשבון את המרחק בין הביטויים, ואת מידת הפיזור שלהם – כל זאת על פני הטבלה. כדי להבהיר זאת, נשוב ונתבונן בטבלה 1 מן הפרק הקודם:

טבלה ט-1



חץ אחד מסמן את המרחק הקצר ביותר בין המד"ש המינימלי[1] של המלה "הא/להים" למד"ש המינימלי[2] של המלה "בוראכם". חץ שני מסמן את המרחק בין שתי אותיות עוקבות של המד"ש "בוראכם". חץ שלישי מסמן את המרחק בין שתי אותיות עוקבות של המד"ש "הא/להים". כל המרחקים נמדדים על פני הטבלה. למשל, המרחק בין שתי אותיות עוקבות של המלה "הא/להים", הוא אות אחת בדיוק. הכימות יעדיף מפגשים שבהם שלושת המרחקים הללו הנם קטנים. כלומר, המפגש – "דחוס".

נספח א2 מתאר כיצד נבנתה "מידת הקִרבה" על פי עקרונות אלה בדרך פשוטה וטבעית. באמצעות "מידת הקִרבה" של זוג מד"שים אנו מקבלים ערך מספרי בשביל כל מפגש בין צמד מד"שים. ככל שהמד"שים מינימליים "יותר", קרובים "יותר" ומכונסים "יותר" – כלומר, המפגש מוצלח "יותר" – יהיה הערך המספרי של "מידת הקִרבה" גבוה יותר.

כל שתיארתי עד כה, אינו אלא הצעד הראשון בדרך לחישוב הסתברות המפגשים. עד כאן תיארתי בצורה איכותית את בניית "מידת הקִרבה" עבור מפגש מסויים בין שני מד"שים, כי היא

---

[1] מינימלי בקטע שאורכו 3/4 מספר בראשית.
[2] מינימלי על כל ספר בראשית.

המשקפת את ההתרשמות האינטואיטיבית שלנו מן המפגשים שכבר ראינו, ואף עתידים אנו לראות, בספר זה. אין בדעתי להאריך כאן בעניינים טכניים, שמקומם בנספח. נספח א2 עונה על השאלות הבאות:

- כיצד מגדירים "מידת קִרבה" בין ביטויים?
- כיצד מעריכים את הסתברות המפגש בין שני ביטויים?

כפתרון לשתי השאלות הללו פותחה "מידת הקִרבה המכוילת" (המוגדרת בנספח א2). זו היא פונקציה המושתתת על "מידת הקִרבה" הנזכרת לעיל. "מידת הקִרבה המכוילת" בין שני ביטויים זה מספר בין 0 ל-1. ככל שמופעי הביטויים כצפני *ELS* מינימליים בקטעים גדולים יותר, וככל שהם קרובים יותר זה לזה ופחות מפוזרים על פני הטבלאות, אזי "מידת הקִרבה המכוילת" קטנה וקרובה יותר ל-0. לעומת זאת "מידת הקִרבה המכוילת" של זוג ביטויים קרובה ל-1 כאשר צפני ה-*ELS* שלהם רחוקים ומפוזרים במיוחד. מידה זו היא הסתברות המפגשים בין שני ביטויים.

לא נותר אלא להפוך פתרונות אלה לתוכנת מחשב. יואב רוזנברג התנדב להתמודד עם משימה קשה זו. היה עליו להקדיש זמן רב כדי לבנות תוכנה יעילה, המתחשבת במגבלות החמורות של אמצעי החישוב שעמדו לרשותנו – והוא עשה זאת בהצלחה. בסוף שנת התשמ"ה (סתיו 1985 למניינם) היתה בידינו תוכנה, המסוגלת למדוד את המפגשים השייכים לתכונה א: מפגשים בין צפני *ELS* לבין עצמם. אתה מקליד למחשב צמד ביטויים, וכפלט מתקבל מספר שהוא ערכה של "מידת הקִרבה המכוילת". כאמור, מספר זה מודד את ההסתברות למפגשים של זוג הביטויים כצפני *ELS*.

--- --- ---

<u>הערה</u>: יש להבהיר, בהנחה שהביטויים אכן מופיעים בספר בראשית בדילוג שווה של אותיות, תמדוד התוכנה מהו הסיכוי שהמפגש יהיה גם "מינימלי" וגם "דחוס". אך התוכנה איננה מודדת מהו הסיכוי, שהביטויים יופיעו בדילוג שווה בבראשית. ואם יתמה הקורא מדוע איננו מתחשבים בכך, אזכיר עקרון ידוע: הגישה הנכונה לחקר תופעה נעלמה היא דווקא לבודד תכונות מסוימות. הנטייה למפגשים היא תכונה אחת. הופעה לא-מסתברת של ביטוי ארוך ו/או מורכב מאותיות נדירות כמד"ש – היא תכונה אחרת. יתכן מאד, שכדאי לחקור גם תכונה זאת – אך בשלב זה, חשוב יותר להפריד בין התכונות.