ב״ה

מדידת מפגשים

מאת דורון ויצטום

כאן נתאר:

- כיצד מגדירים ״מידת קרבה״ בין ביטויים.
- כיצד מעריכים את ההסתברות של מפגש בין שני ביטויים.

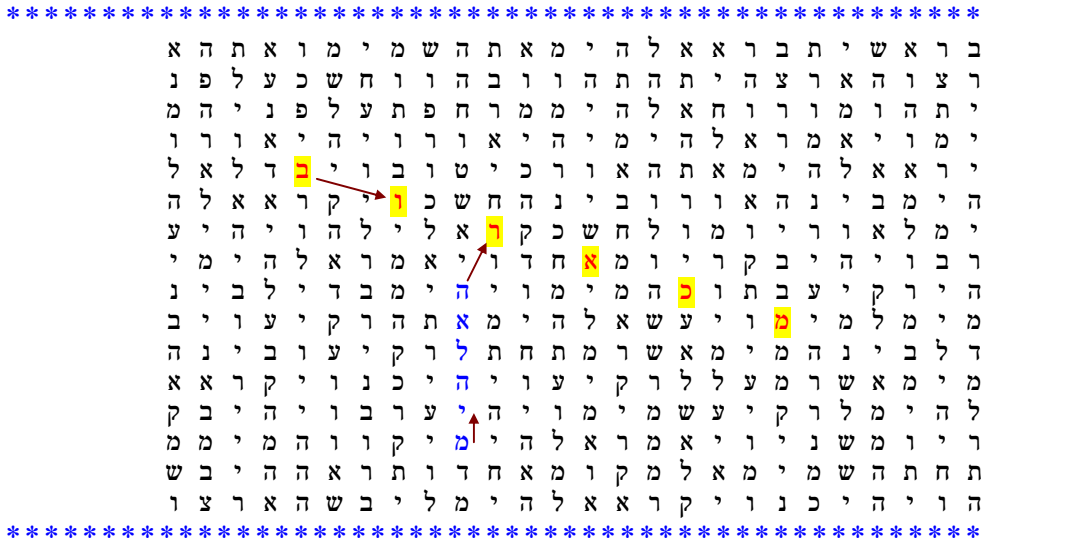
כל זה ייעשה (בצורה דומה) הן עבור מפגשים מטיפוס א׳ (מפגש בין מד״שים) בסעיפים א׳-ב׳, והן עבור מפגשים מטיפוס ב׳ (מפגשים בין מד״שים לבין ביטויים הנמצאים כרצף אותיות בטקסט) – בסעיף ג׳ להלן. בסעיף ד׳ אציג לראשונה יישום של הרעיונות המקוריים, המאפשר לחשב את ההסתברות למפגש נתון על פני טבלה מסוימת, ובעקבות זאת לנקוב בסיכוי למפגשים כגון אלה המוצגים באתר.

בנספח ״טעות פשוטה״ ישנה תוספת ביאור והדגמה של כמה מן הרעיונות, הניתנים כאן בצורה פורמלית יותר.

א. מידת הקרבה בין ביטויים: **המקרה של תכונה א׳.**

מפגש בין ביטויים הוגדר על ידינו כ״סכום״ המפגשים של המד״שים המייצגים את הביטויים. לכן עלינו להתחיל ממפגש בין שני מד״שים בודדים.

מאפיין אחד של המפגש הוא **טיב המפגש**: מפגש ״טוב״ נראה כמקבץ מכונס; כלומר, המד״שים מופיעים על פני הטבלה, כאשר הם קרובים זה לזה ואינם מפוזרים. אנו נרצה למצוא מדד מתאים ל**טיב המפגש.** מדד זה יעניק ציון ״טוב יותר״ למפגשים שבהם מופיעים המד״שים כשהם קרובים זה לזה ואינם מפוזרים. זה המדד ל״דחיסות״ המפגש. לפי גישתו של פרופסור אליהו ריפס, המדד ל״דחיסות״ לוקח בחשבון את המרחק שבין הביטויים ואת מידת הפיזור שלהם – **כל זאת על פני הטבלה**. נתבונן בטבלה הבאה:



הבה נמדוד את כל המרחקים על פני הטבלה. למשל, המרחק בין שתי אותיות עוקבות של המלה "הא/להים" (מסומן בחץ קטן), הוא אות אחת בדיוק. המרחק בין שתי אותיות עוקבות של המלה "בוראכם" (מסומן בחץ העליון), הוא $\sqrt{1^2+3^2}=\sqrt{10}$ (משפט פיתגורס!). ואילו המרחק הקצר ביותר בין "הא/להים" ל"בוראכם" (המסומן בחץ השלישי), הוא $\sqrt{1^2+2^2}=\sqrt{5}$.

עתה נתאר באופן כללי את שעשינו בדוגמא זו.

אנו מתחילים בטקסט המוצג בצורה דו-ממדית: על פני גליל שהיקפו h – כלומר, h הוא מספר הטורים במאונך (בדוגמא לעיל h=26). כל טבלה שקולה לגליל, ולהיפך.

כדי להגדיר את המרחק בין האות שמספרה x (מתחילת הטקסט) לאות שמספרה x', נחתוך את הגליל לאורך קו אנכי בין שני טורים. כך מקבלים טבלה מישורית, שלכל אות בה שתי קואורדינטות במספרים שלמים. נחשב את המרחק בין x ל- x' באמצעות הקואורדינטות שלהן, כמקובל. (בדרך כלל יהיו שני ערכים אפשריים למרחק זה, בהתאם למיקומו של חיתוך הקו האנכי. במקרה ששני הערכים שונים, נשתמש בערך הקטן.)

עתה נתאר את טיב המפגש בין מד"ש מסוים, e, לבין מד"ש מסוים אחר, e'. (בדוגמא שלפנינו, e הוא המד"ש של "הא/להים", ו- e' הוא המד"ש של "בוראכם").

נסמן ב- f את המרחק בין שתי אותיות עוקבות של e, וב- f' את המרחק בין שתי אותיות עוקבות של e'. כן נסמן ב- l את המרחק המינימלי בין אות של e לאות של e'.

(בדוגמא שלנו, $f=1$, $f'=\sqrt{1^2+3^2}=\sqrt{10}$, $l=\sqrt{1^2+2^2}=\sqrt{5}$.)

נגדיר:

$$\mu_h(e,e')\equiv\frac{1}{f^2+f'^2+l^2} \qquad (3.1)$$

מן ההגדרה נובע שהפונקציה $\mu_h(e,e')$ מקבלת ערך מספרי גבוה אם, ורק אם, המפגש **מכונס**: כלומר, f, f' ו- l הם קטנים. פונקציה זו, המודדת את "דחיסות" המפגש על הטבלה, תעניק ציון "טוב יותר" למפגשים שבהם מופיעים המד"שים כשהם קרובים זה לזה ולא מפוזרים, על פני הטבלה.

(בדוגמא שלנו, $\mu_h(e,e')=\frac{1}{1+10+5}=\frac{1}{16}$.)

היקף הגליל h אינו שרירותי: הוא נקבע על ידי המד"שים e ו- e' עצמם. דהיינו, אם d הוא הדילוג של e, אנו מתעניינים בגלילים שהיקפם הוא הערך השלם של:

$$h_1=|d|,\ h_2=|d|/2,\ h_3=|d|/3,\ldots,\ h_{10}=|d|/10 \qquad (3.2)$$

(1/2 מעוגל למעלה). ובדומה, אם d' הוא הדילוג של e', אנו מתעניינים בגלילים שהיקפם הוא הערך השלם של:

$$h'_1=|d'|,\ h'_2=|d'|/2,\ h'_3=|d'|/3,\ldots,\ h'_{10}=|d'|/10 \qquad (3.3)$$

טיב המפגש של שני המד"שים e ו- e' יוגדר כסך כל "דחיסות" המפגשים על כל הגלילים:

$$\sigma(e,e')\equiv\sum_{i=1}^{10}\mu_{h_i}(e,e')+\sum_{i=1}^{10}\mu_{h'_i}(e,e') \qquad (3.4)$$

כאשר $h_i=\lfloor|d|/i\rfloor$, $h'_i=\lfloor|d'|/i\rfloor$.

טיב המפגש, שהוא הפונקציה $\sigma(e,e')$, מקבל ערך גבוה אם, ורק אם, ישנו מפגש **מכונס** בין שני המד"שים e ו- e', לפחות על אחד הגלילים הנ"ל.

החישוב של **טיב המפגש** דומה עקרונית לדוגמא שלפנינו. אלא שיש לחזור על החישובים לגבי כל גליל (טבלה) משתי סדרות הגלילים (הטבלאות) שקבעו המד"שים, ואחר כך לסכם את התוצאה.

המאפיין השני של המפגש הוא **טיב הנפגשים**. אנו מתעניינים במפגש המתקיים בין מד"שים, שהם מינימליים בקטעים גדולים בספר. **טיב הנפגשים** יתואר על ידי הפונקציה $\omega(e,e')$, המבטאת את ההעדפה של ההופעות המינימליות.

נגדיר את תחום המינימליות, T_e, של המד"ש e, כקטע הארוך ביותר של הטקסט המכיל את e, אך אינו מכיל מד"ש בעל דילוג קצר ממנו. בדומה נגדיר את $T_{e'}$. לכן, הקטע הארוך ביותר בטקסט, בו **שני** המד"שים הללו הם מינימליים, הוא חיתוך התחומים: $T_e \cap T_{e'}$. הוא ייקרא בשם "תחום המינימליות המשותפת של e ו- e'".

הפונקציה $\omega(e,e')$ מוגדרת כיחס בין אורך הקטע $T_e \cap T_{e'}$ לאורך הטקסט כולו. [אם הקטע $T_e \cap T_{e'}$ ריק, אזי $\omega(e,e')=0$.]

(בדוגמא שלפנינו:

T_e, הקטע הארוך ביותר של ספר בראשית המכיל את המד"ש e של המלה "הא/להים", אך אינו מכיל מד"ש של מלה זו בדילוג קצר ממנו, מתחיל באות הראשונה בספר בראשית ומשתרע עד האות מס' 57,319. כלומר: $T_e=[1,57319]$.

$T_{e'}$, הקטע הארוך ביותר של ספר בראשית המכיל את המד"ש e' של מלה "בוראכם", ואינו מכיל מד"ש של מלה זו בדילוג קצר ממנו, מתחיל באות הראשונה בספר בראשית ומשתרע עד האות האחרונה שהיא אות מס' 78,064. כלומר: $T_{e'}=[1,78064]$.

במקרה זה, תחום המינימליות המשותפת יהיה הקטע המתחיל באות הראשונה ומסתיים באות מס' 57,319. כלומר: $T_e \cap T_{e'}=[1,57319]$.

לכן, המדד של טיב הנפגשים יהיה בדוגמא זו: $\omega(e,e')=\frac{57319}{78064}=0.73$.)

מעתה, יכולים אנו להגדיר את "**מידת הקִרבה**" של צמד המד"שים e ו- e'. זו תוגדר כמכפלה $\omega(e,e')\sigma(e,e')$, שהיא מכפלת **טיב המפגש בטיב הנפגשים**.

מהגדרת טיב הנפגשים וטיב המפגש נובע, כי "מידת הקִרבה" של צמד מד"שים תקבל "ציון גבוה" ככל שהמד"שים יהיו מינימליים "יותר", קרובים "יותר" ומכונסים "יותר", לפחות על אחד מן הגלילים המוגדרים ב- (3.2-3.3).

מפגש בין ביטויים הוגדר על ידינו כ"סכום" המפגשים של המד"שים המייצגים את הביטויים. על כן, לאחר שהגדרנו את "**מידת הקִרבה**" של זוג מד"שים, נוכל לעבור סוף סוף להגדרה של **"מידת הקִרבה" של צמד ביטויים.**

"מידת הקִרבה" של צמד ביטויים w ו- w' תוגדר כסכום:

$$\Omega(w,w') \equiv \sum \omega(e,e')\sigma(e,e') \qquad (3.5)$$

בסכום זה אנו מסכמים את כל המד"שים e ו- e' של הביטויים w ו- w'.

יש להדגיש, שלפי גישת "המסנן" שנקטנו (ראה ב"צופן בראשית", סוף פרק ב'), אנו מעונינים לעקוב אחר המד"שים שהם מינימליים בקטעים גדולים בספר בראשית. לכן, בדקנו את הופעות הביטוי w כמד"שים רק עד דילוגים בגודל $D(w)$. גודל זה חושב כך, שתוחלת מספר המד"שים של w עד דילוג זה אינה עולה על 10. כך נהגנו גם לגבי המד"שים של הביטוי w', הם נבדקו עד לדילוגים בגודל $D(w')$. ולכן ייעשה הסיכום בהגדרת $\Omega(w,w')$ רק עבור המד"שים של w ו- w' בתחומי הדילוגים הללו.

"מידת הקרבה" שהגדרנו, משקפת את תכונות המפגשים בין המד"שים של הביטויים w ו- w': ככל שהמפגשים יהיו קרובים ומכונסים יותר ואילו המד"שים יהיו מינימליים בקטעים גדולים יותר של טקסט – כך תגדל "מידת הקרבה" $\Omega(w,w')$.

ב. מידת הקרבה המכוילת: **המקרה של תכונה א'.**

מעתה מסוגלים אנו לחשב מה ערכו המספרי של מפגש של צמד ביטויים. אך מה משמעותו של המספר שקבלנו? האם זה הערך שצפויים היינו לקבל במקרה? או שמא גבוה ממנו, או נמוך ממנו? מה הסיכוי לקבל ערך כזה באקראי?

אילו ידעתי לחשב תיאורטית את התפלגות ערכי "מידת הקרבה" של המפגש, הצפויים לעלות באקראי, הרי היתה בידי תשובה לכל השאלות הללו... אבל חישוב כזה הנו מסובך מדי, וכלל לא מעשי.

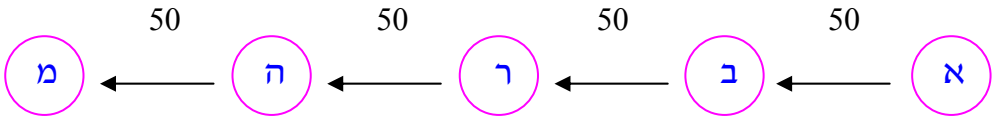
הפתרון המעשי שהציע פרופסור אליהו ריפס, מבוסס על תהליך מקובל של השוואה (שיטת מונטה-קרלו), תהליך שאפשר לתארו באופן ציורי כך:

נתאר לעצמנו שעורכים תחרות בין "רצים" שווי "כוחות": כלומר, לכל "רץ" יש אותו סיכוי להגיע לכל מקום בדירוג "הרצים". במצב כזה אנו יודעים לחשב את הסיכוי של "רץ" להגיע עד מקום מסוים בדירוג. למשל, אם 100 "רצים" משתתפים בתחרות, אזי יש ל"רץ" סיכוי של 1/10 להגיע לאחד מעשרת המקומות הראשונים.

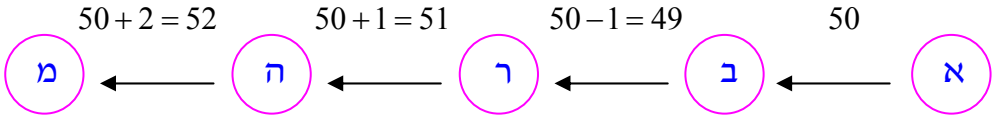
אם נוכל לערוך "מרוץ" דומה בין "מידת הקרבה" של צמד ביטויים לבין "רצים" אחרים, הרי פתרנו את הבעיה... אבל **מי** יהיו "הרצים" האחרים?

וכאן עיקר רעיונו של ריפס: הוא הציע להשתמש בהשוואה בין המד"שים לעומת הופעות של **מ**לים ב**ד**ילוגים **כ**מעט **ש**ווים – **מדכ"שים**. המדכ"שים, כפי שניווכח מיד, הם מד"שים שעברו מוטציה שֶׁשיבשה את גודל הדילוג שלהם.

עד כה עסקנו בהופעת מלים בדילוג שווה של אותיות. למשל, המלה "אברהם" בדילוג שווה של 50 תיראה כך:

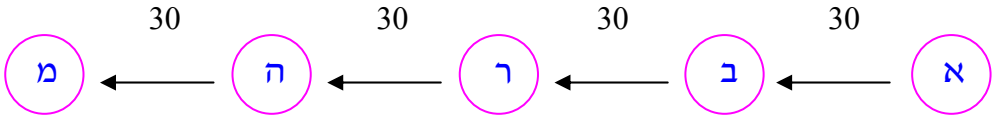


נתבונן עתה בהופעתה של מלה זו בדילוג משובש קמעה, דילוג שהוא **כמעט** שווה:

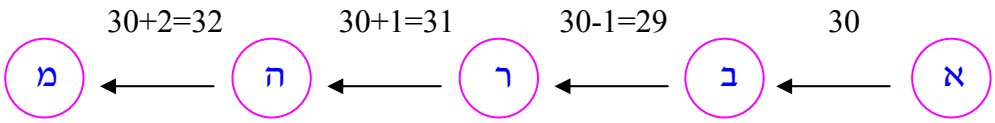


האות "ב" היא האות ה-50 מן האות "א"; האות "ר" – האות ה-49 מן האות "ב"; האות "ה" היא האות ה-51 מן האות "ר"; והאות "מ" – האות ה-52 מן האות "ה".

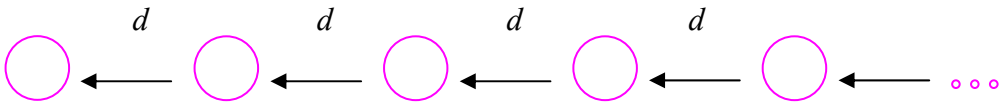
ובדומה: במקביל להופעת המלה "אברהם" בדילוג שווה של 30:



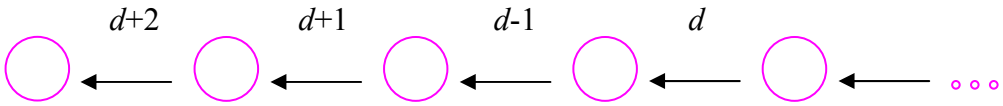
אפשר לחפש אחר הופעת מלה זו בדילוג המשובש (כמעט שווה):



ובאופן כללי: במקביל להופעת ביטוי כלשהו בדילוג שווה של d:

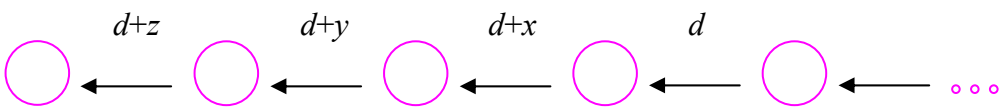


אפשר להתבונן בהופעת ביטוי זה בדילוג המשובש (כמעט שווה):



בדוגמא זו, מגדירים שלושת המספרים (-1, 1, 2) סוג מסוים של שיבוש. כך אנו יכולים לחפש אחר **מדכ"שים** - מלים בדילוגים כמעט שווים – לפי חוקיות קבועה זו, כאשר d יכול לקבל כל ערך שהוא, בדיוק כשם שאנו מחפשים אחר **מד"שים** – ביטויים בדילוגים שווים של אותיות – כאשר d יכול לקבל כל ערך שהוא (במגבלות של אורך הטקסט). בפרט אנו יכולים לאתר את המדכ"שים המינימליים, לבטא את המפגשים ביניהם על פני טבלאות דו-ממדיות, ולהגדיר "מידת קרבה", $\Omega^{(-1,1,2)}(w,w')$, עבור המלים w ו-w', באמצעות מדכ"שים אלה – בדיוק באותו האופן שבו הוגדרה "מידת הקרבה" $\Omega(w,w')$ עבור אותן המלים באמצעות המד"שים.

שלושת המספרים (-1,1,2) שימשו בדוגמא זו להגדרת סוג מסוים של שיבוש; שְׁלָשָׁה אחרת של מספרים שלמים תגדיר סוג אחר של שיבוש. ובאופן כללי: כל שְׁלָשָׁה (x,y,z) של מספרים שלמים תגדיר סוג מסוים של שיבוש:



כיוצא בזה, עבור כל סוג של שיבוש ניתן להגדיר "מידת קירבה" $\Omega^{(x,y,z)}(w,w')$ של צמד המלים w ו- w' באמצעות מדכ"שים אלה.

עכשיו, סוף סוף, אפשר לערוך תחרות: בהינתן זוג ביטויים, יחפש המחשב אחר **המד"שים** שלהם, שהם מינימליים בקטעים גדולים בספר, ויחשב את "מידת הקרבה" $\Omega(w,w')$ באמצעות מפגשיהם על פני הגלילים (הטבלאות) הנקבעים על ידם. כיוצא בזה, **באותו האופן עצמו,** יחפש המחשב עבור השיבוש (x,y,z) אחר **המדכ"שים** שלהם, שהם מינימליים בקטעים גדולים בספר, ויחשב את "מידת הקרבה" $\Omega^{(x,y,z)}(w,w')$ באמצעות מפגשיהם על פני הגלילים (טבלאות) הנקבעים על ידם. אם עושים זאת לגבי קבוצה גדולה של סוגי שיבושים – מקבלים קבוצה גדולה של "רצים": זו קבוצת ערכי "מידת הקרבה" בשיבושים השונים.

אם אכן הופעת מד"שים ומדכ"שים בטקסט אינה אלא **אקראית**, הרי אלו "רצים" שווי "כוחות". המחשב יערוך "מירוץ" בין "הרצים" הללו, וידרג אותם: במקום הראשון תעמוד "מידת הקרבה" הגדולה ביותר, אחריה "מידת הקרבה" הגדולה משאר המתחרים, וכו'. אנו מתעניינים אפוא בדירוג של "מידת הקרבה" של המד"שים. ליתר דיוק, אנו רוצים לדעת: מה ההסתברות, שהדירוג של "מידת הקרבה" הזאת יהיה כזה או גבוה יותר?

זאת קל לדעת: כל מה שיש לעשות הוא – לחלק את המספר המבטא את הדירוג במספר המשתתפים "במרוץ". (למשל, אם בין 125 מתחרים דורג הרץ "שלנו" – "מידת הקרבה" של המד"שים" – במקום ה-5, הסיכוי לקבל דירוג כזה או גבוה ממנו הוא: $p=5/125=0.04$.)

נותר לנו רק לתאר את התחרות בצורה מדויקת.

אם שלושת המספרים השלמים (x,y,z) מצויים בטווח הערכים $\{-2,-1,0,1,2\}$, הרי ישנן $5\times5\times5=125$ שלשות כאלו – כלומר 125 סוגי שיבושים. שים לב, כי אחת מן השלשות היא השלשה $(0,0,0)$ – כלומר, הדילוג שווה (אין שיבוש). עבור 125 סוגי השיבושים, נקבל 125 מספרים $\Omega^{(x,y,z)}(w,w')$, **שאחד מהם** הוא "מידת הקרבה" בדילוגים השווים: $\Omega(w,w')=\Omega^{(0,0,0)}(w,w')$. נסדר את המספרים $\Omega^{(x,y,z)}(w,w')$ לפי גודלם, ונסמן ב- $v(w,w')$ את המספר הסידורי (הדירוג) של $\Omega(w,w')$.

יתכן שקיימות שלָשות (x,y,z), שהביטויים w ו/או w' אינם מופיעים בדילוג המשובש לפי (x,y,z). נסמן ב-$m(w,w')$ את מספר השלָשות שבהן מופיעים גם w וגם w'.

נגדיר את **"מידת הקרבה המכוילת"**:

$$c(w,w') \equiv v(w,w')/m(w,w') \qquad (3.6)$$

"מידת הקרבה המכוילת" היא, אם כן, הדירוג של "מידת הקרבה" בדילוג השווה "במרוץ" בין "מידות הקרבה", כשהוא מחולק בכלל משתתפי "המרוץ". היא מהווה אפוא מדד להסתברות, שערכה של "מידת הקרבה" הוא כה גבוה.

ערכי $c(w,w')$ הם בין 0 ל- 1: הערך קרוב ל- 0 ("הרץ שלנו" קרוב לראש הדירוג), כאשר מד"שים מינימליים של הביטויים נפגשו בצורה מכונסת במיוחד. הערך קרוב ל- 1 ("הרץ שלנו" מזדנב בסוף), כאשר המד"שים הנ"ל רחוקים ומפוזרים במיוחד...

יש לשים לב להגבלות הבאות:

1. $c(w,w')$ מוגדרת אך ורק כאשר גם w וגם w' מופיעים בדילוג שווה.
2. כאשר $m(w,w')$ (מספר "הרצים") קטן מ- 10, לא הגדרנו את "מידת הקרבה המכוילת", כי לדעתנו המדידה אינה מדויקת די הצורך: אין די "מתחרים". [דוגמא קיצונית: נניח שהמירוץ נערך לגבי זוג ביטויים שבהם אותיות נדירות, כך שלמעשה רק שני "רצים" מופיעים על המסלול: המד"שים, ועוד אחד מן המדכ"שים (כלומר: $m(w,w')=2$). עוד נניח, שהמד"שים ניצחו "במירוץ" ($v(w,w')=1$). במקרה זה נקבל לפי הגדרה (3.6), כי $c(w,w')=1/2=0.5$. תוצאה זו, שאינה קרובה ל - 0, מצביעה על כישלון!]
3. הגדרת "מידת הקרבה המכוילת" מבוססת על השוואה לשיבושים (x,y,z), ולכן ניתן ליישם אותה בביטויים בני 5 אותיות לפחות. זאת ועוד, כאשר משתמשים בהשוואה ל- 125 השיבושים (x, y, z), עבור ביטויים w או w' שבהם למעלה מ- 8 אותיות, הרי $m(w,w')$ (מספר "הרצים") *קטן כמעט תמיד* מ- 10. לכן יישמנו מידה זו רק בביטויים בני 5-8 אותיות.

הגבלות אלה בהגדרות המקוריות שלנו באו, כמובן, לכלל ביטוי בתוכנה ששימשה בניסוי הגדול הראשון. מאוחר יותר, מצאנו דרך להסיר חלקית את ההגבלה השלישית (ההגבלה לביטויים בני 5-8 אותיות). על כך – ב"צופן בראשית", בנספח על מדגמי "כותרת". מומלץ מאוד לקרוא בקובץ "טעות פשוטה" על **ההכרח** לכייל את "מידת הקרבה".

ג. מידת הקרבה המכוילת: **המקרה של תכונה ב'.**

כאן נגדיר את "מידת הקרבה המכוילת" עבור מפגשים מטיפוס ב' של זוג ביטויים (כלומר, מפגשים בין מד"שים מינימליים לבין ביטויים הבאים כרצף אותיות בטקסט). למעשה, כל הרעיונות ששימשו אותנו בהגדרת "מידת הקרבה המכוילת" לגבי תכונה א', ניתנים ליישום במקרה של תכונה ב'. לא נותר לנו אלא להכניס שינויים קלים המתחייבים מן השוני בטיפוסי המפגשים.

1. בזוג הביטויים (w,w') – רק w נבדק בדילוג השווה, בעוד ש- w' נבחן כרצף אותיות בטקסט, כלומר, ב"דילוג" 1 או 1–. לכן, הגלילים הנבדקים נקבעים רק על ידי המד"שים של w. כתוצאה מכך, בהגדרת $\sigma(e,e')$ [ראה (3.4) לעיל] נסכם רק על h_i:

$$\sigma(e,e') \equiv \sum_{i=1}^{10} \mu_{h_i}(e,e') \qquad (3.7)$$

2. כאשר מחשבים את $\Omega^{(x,y,z)}(w,w')$, בודקים רק את המדכ"שים של w בשיבושים (x,y,z), ואילו ההופעות של w', הן אותן ההופעות e' כרצף אותיות $(d=\pm 1)$, בדומה למפגשים עם המד"שים.

3. כיוון שהשיבושים (x,y,z) אינם מעורבים בהופעות של w', **אין צורך להגביל את אורך הביטוי w'.**

4. בהגדרת הדירוג - $v(w,w')$ - מביאים בחשבון מקרי "תיקו": אם ישנם "רצים", שערך מידותיהם שווה לערך מידת ה"רץ שלנו" (כלומר, מקרה של "תיקו"), ייחשבו חציים כמקדימים את ה"רץ שלנו" בדירוג. [תיאורטית – כך צריך להיעשות גם בסעיף ב' לעיל ובסעיף ד' להלן. אך מעשית, השינוי בהגדרה אינו משמעותי עבור תכונה א', שבה מקרי "תיקו" נדירים.]

ד. מידת הקרבה המכוילת: **מרוץ של "אלופים".**

בסעיף א', התחלנו מן ההגדרה של "מפגש בין ביטויים" כ"**סכום**" המפגשים של המד"שים המייצגים אותם. נקודת מוצא זו היא שהובילה להגדרת "מידת הקרבה" בין ביטויים, כסכום הערכים המספריים של "מידת הקרבה" של המד"שים המייצגים אותם (ראה הגדרות (3.5) ו-(3.7) לעיל).

למעשה, קיימת אפשרות אחרת להגדרת "מידת הקרבה" של ביטויים: להגדיר אותה כערך המספרי של **המפגש המסויים** "המוצלח ביותר" בין המד"שים המינימליים המייצגים את הביטויים. אם נשתמש במינוח של סעיף א', הרי במקום הסכום

$$\Omega(w,w') \equiv \sum \omega(e,e')\sigma(e,e') \qquad (3.5)$$

בו אנו מסכמים את כל הערכים המספריים של "מידת הקרבה" של המד"שים e ו- e' של הביטויים w ו- w' (בתחומי הדילוגים הנתונים), כאן, לעומת זאת, לוקחים אך ורק את ערכו המספרי של המפגש "האלוף":

$$\Omega_B(w,w') = \max\{\mu_{h_i}(e,e')\omega(e,e')\} \qquad (3.8)$$

כאשר ערך זה הוא המכסימלי בקבוצת כל המפגשים על הגלילים h_i ו- $h_{i'}$ של כל המד"שים e ו- e' של הביטויים w ו- w' (בתחומי הדילוגים הנתונים). לפי הגדרה חלופית זו, "מידת הקרבה" של הביטויים היא הערך המספרי של המפגש "האלוף" – המפגש המוצלח ביותר – שקיים עבורם על אחד הגלילים (הטבלאות). נסמן אפוא "מידת קרבה" זו ב- BEST.

פרופ' ריפס העדיף את שיטת "הסכום" (המוגדרת ב-(3.5)), בגלל שיקולים תיאורטיים של יציבות. אני העדפתי את שיטת "הסכום", משום שהתרשמותי היתה (ועודנה) כי ההצפנה של נושא בטקסט הנסתר מבוססת במקרים רבים על מפגשים **חוזרים**, ולא על מפגש יחיד: כמה מד"שים של "מלה א" נפגשים עם כמה מד"שים (ו/או עם רצף אותיות) של "מלה ב". ואכן, המחקר שנערך עד כה התנהל לפי שיטת "הסכום".

אבל, יש מקום לשימוש בשיטת BEST. אם מטרתנו היא להעריך את ההסתברות של מפגשים מסויימים על פני טבלאות נתונות, כגון אלו המופיעות באתר

זה, נראה שהדרך הנאותה לכך היא באמצעות שיטת BEST. לצורך זה, נגדיר את "מידת הקרבה המכוילת" בשיטה זו. הדבר נעשה על ידי תחרות של המד"שים עם המדכ"שים. הפעם התחרות היא בין המפגש "האלוף" של המד"שים לבין המפגשים "האלופים" של המדכ"שים. גם כאן, נדרג את "הרצים", ונקבע את הדירוג – $v_B(w,w')$ - של "אלוף" המד"שים, מבין $m_B(w,w')$ "הרצים" המתחרים, שהם "אלופי" המדכ"שים. כך נקבל את "מידת הקרבה המכוילת" בשיטת BEST:

$$c_B(w,w') \equiv v_B(w,w')/m_B(w,w') \qquad (3.9)$$

(השווה (3.6)). זהו בעצם מדד להסתברות, שערכו המספרי של המפגש "האלוף" של המד"שים יהיה כה גבוה.