הערכת רמות קושי של טקסטים המשמשים בהערכת יכולת הבנת הנקרא, באמצעות מאפיינים סטטיסטיים ומורפולוגיים של הטקסט

דנה רובינשטיין, יעל שפרן, ענת בן-סימון

פרויקט השפה העברית (HLP), מאל"ו

הכנס השביעי של אפ"י, 2011, ירושלים

רקע

- מחקר זה עוסק באיתור מדדים להערכת קושי של טקסטים המשמשים בבחינות שונות של מאל"ו להערכת יכולת הבנת הנקרא
- מחקרים בנושא ניבוי קושי של טקסטים מוכרים בספרות המקצועית כמחקרי "**קריאות**" (readability)

הגדרת קריאות (readability)

- “The ease of understanding or comprehension due to the style of writing” (George Klare, 1963)

- בשפה האנגלית פותחו נוסחאות קריאות המשמשות לצרכים שונים

קריאות (readability), דוגמה

"Natalie Portman has a big year ahead of her – a new engagement, a baby on the way and a big awards season coming up…"

(People Magazine)

"The histrionics are there precisely in Suleiman's understatement; the film's polemical and rhetorical edge…"

(The New Yorker)

69

Flesch Reading Ease Formula (0-100)

48

www.readabilityformulas.com

מדדי קריאות

המשתנים הרווחים בנוסחאות קריאות בשפה האנגלית הם:

- **מאפיין של מורכבות תחבירית**
  למשל:
  - אורך משפט ממוצע
- **מאפיין של רמה לקסיקלית**
  למשל:
  - ממוצע של מספר ההברות במילה
  - אחוז המלים ה"קשות"
  - אורך מילה ממוצע

המחקר

- **<u>מטרת המחקר:</u>** זיהוי משתנים המנבאים קושי בטקסטים המשמשים להערכת יכולת הבנת הנקרא, ופיתוח נוסחת קריאות לניבוי קושי של טקסטים מסוג זה.
- **<u>השערה:</u>** נקבל מאפיינים למורכבות תחבירית ולרמה לקסיקלית, בדומה למשתנים המשמשים בנוסחאות הקריאות הקיימות.

שיטת המחקר: כלים

טקסטים

- 70 קטעים המשמשים להערכת יכולת הבנת הנקרא בבחינות שונות של מאל"ו
- הטקסטים עוסקים בתחומי-דעת מגוונים (ביולוגיה, פילוסופיה, פסיכולוגיה, היסטוריה, וכו')
- הטקסטים מופיעים במבחן כשהם מלווים במספר שאלות מסוג בררה הבוחנות את הבנתם

מאפייני טקסט

- 70 מאפייני טקסט שהופקו על-ידי הNITE-Rater

שיטה: הליך

- ההגדרה האופרציונלית של קושי הטקסט:
  שיפוט של מומחים מצוות הפיתוח
- 70 הטקסטים חולקו ל-7 מקבצים של 10 טקסטים
- כל מקבץ הועבר לשיפוט של 3 שופטים
- השיפוט בוצע על סולם 1-10, תוך שימוש ב-3 טקסטים לעיגון ההערכות
- *"דרג את הטקסטים מהקל אל הקשה, ומקם אותם על הציר [1-10] תוך התייחסות לקטעי העוגן. ניתן להעריך מספר טקסטים כבעלי רמת קושי זהה."*
- לכל טקסט חושב ציון קושי ממוצע

תוצאות

מקבץ	ממוצע המתאמים בין זוגות שופטים
1	.64
2	.76
3	.76
4	.37
5	.71
6	.74
7	.53

מידת ההסכמה בין השופטים בנוגע לקושי הטקסטים

תוצאות

- ל-20 מתוך 70 המאפיינים נמצא קשר מובהק (r>.23) עם הערכת רמת הקושי של הטקסטים.
- את המאפיינים המשמעותיים חילקנו לשלוש קטגוריות:
  1. מאפיינים של מורכבות תחבירית
  2. מאפיינים של רמה לקסיקלית
  3. מאפיינים נוספים

מתאמים בין מאפיינים לבין קושי טקסט

מאפיין	תיאור המאפיין	מתאם
שיעור מילות התוכן	שמות עצם, תואר, פעלים...	-.46
מספר המלים הכולל		.38
סטיית התקן של אורכי משפטים		.37
שיעור השלילות	לא, אין	.31
שיעור מילות השעבוד	ש-, אשר, כיוון ש-	.29
שיעור המלים הנדירות ביותר		.28
שיעור הכינויים	אני, הוא, זה	.26
גיוון לקסיקלי		-.26

תוצאות: מאפיינים של מורכבות תחבירית

מאפיין	תיאור המאפיין	מתאם
שיעור מילות התוכן	שמות עצם, תארים, שמות פרטיים, פעלים, מלים לועזיות	-.46
שיעור מילות שלילה	לא, אין	.31
שיעור מילות שעבוד	ש-, אשר, כאשר, אחרי ש-, למרות ש-...	.29
שיעור הכינויים	אני, הוא, זה,	.26
שיעור מילות השיח	כלומר, לדוגמה,...	.18

מאפיינים של מורכבות תחבירית:

מילות תוכן ומלים דקדוקיות

<u>מילות תוכן:</u>

שמות עצם (*כלב, אהבה*)

שמות תואר (*גבוה*)

פעלים (*אמרתי, ילמד*)

תארי פועל (*לאט, מחר*)

<u>מלים דקדוקיות ("לא תוכן"):</u>

מילות יחס (*בתוך, על*)

מילות שעבוד (*אשר, ש-, כאשר, כיוון ש-*)

מילות קישור (*אלא, אך*)

כינויים (*אני, היא, הם, זה*)

מילות שיח (*לדוגמה*)

דוגמה לקטעים בעלי שיעור שונה של "מילות תוכן"

- "שבעה אנשים נתפסו בידי שלטונות המכס הישראליים כשניסו לבצע הברחה גדולת-ממדים של בשמים יוקרתיים. במהלך חקירתם הם הודו במעשים קודמים של הברחת מוצרי קוסמטיקה לישראל, וחוקרי המכס התרשמו שמדובר בהברחה שיטתית. לדברי מנהל המכס, החשודים נהגו לרכוש את הבשמים ברוסיה ברובלים, למכור אותם בארץ בשקלים, להמיר את השקלים בדולרים, ולצאת מן הארץ כשבכיסיהם כמות גדולה של דולרים."

  (מובאה באורך 57 מילה מתוך קטע בקושי 4.3, עם שיעור מילות תוכן 0.78)

- "אולם נשאלת השאלה - האם אפשר להוכיח את נכונותה של הנחת הסיבתיות? הפילוסוף דייוויד יום טען שמכיוון שעצם קיומו של המדע מבוסס על הנחה זו, אי-אפשר להוכיחה בכלים מדעיים, שכן הוכחה כזו תהיה הוכחה מעגלית, כלומר, הוכחה שמתבססת על ההנחה שמנסים להוכיח. ומה בדבר השאלה ההפוכה - האם אפשר להוכיח את אי-נכונותה של הנחת הסיבתיות?

  (מובאה באורך 55 מילה מתוך קטע בקושי 7.3, עם שיעור מילות תוכן 0.65)

תוצאות: מאפיינים של מורכבות תחבירית

מאפיין	מתאם
סטיית התקן של אורכי המשפט	.37
אורך משפט ממוצע	.12

- "אורך משפט ממוצע" לא מתואם עם קושי הטקסטים.
- יש שונות קטנה מאוד בין הטקסטים עבור משתנה זה.

תוצאות: מאפיינים של רמה לקסיקלית

מאפיין	תיאור המאפיין	מתאם
שיעור המלים הנדירות	שיעור המלים הנדירות לפי קורפוס M1	.28

תוצאות: מאפיינים נוספים

מאפיין	תיאור המאפיין	מתאם
אורך הטקסט	מספר מלים כולל	.38
גיוון לקסיקלי		-.26

- פרשנות אפשרית: העמקה ברעיון

דוגמה לקטעים בעלי רמה שונה של גיוון לקסיקלי

- "שבעה אנשים נתפסו בידי שלטונות **המכס** הישראליים כשניסו לבצע **הברחה** גדולת-ממדים של בשמים יוקרתיים. במהלך חקירתם הם הודו במעשים קודמים של **הברחת** מוצרי קוסמטיקה לישראל, וחוקרי **המכס** התרשמו שמדובר **בהברחה** שיטתית. לדברי מנהל **המכס**, החשודים נהגו לרכוש את הבשמים ברוסיה ברובלים, למכור אותם בארץ בשקלים, להמיר את השקלים בדולרים, ולצאת מן הארץ כשבכיסיהם כמות גדולה של דולרים."

(מובאה באורך 57 מילה מתוך קטע בקושי 4.3, עם ציון גיוון לקסיקלי 7.27)

- "אולם נשאלת השאלה - האם אפשר **להוכיח** את נכונותה של **הנחת** הסיבתיות? הפילוסוף דייוויד יום טען שמכיוון שעצם קיומו של המדע מבוסס על **הנחה** זו, אי-אפשר **להוכיחה** בכלים מדעיים, שכן **הוכחה** כזו תהיה **הוכחה** מעגלית, כלומר, **הוכחה** שמתבססת על **ההנחה** שמנסים **להוכיח**. ומה בדבר השאלה ההפוכה - האם אפשר **להוכיח** את אי-נכונותה של **הנחת** הסיבתיות?

(מובאה באורך 55 מילה מתוך קטע בקושי 7.3, עם ציון גיוון לקסיקלי 7.16)

משוואת רגרסיה לניבוי רמת הקושי

	שם המשתנה	Beta (מקדם מתוקנן)	רמת מובהקות
	אורך הטקסט	.384	.000
	גיוון לקסיקלי	-.309	.003
רמה לקסיקלית	שיעור המלים הנדירות	.305	.001
מורכבות תחבירית	שיעור מילות התוכן	-.271	.006
	ס.ת. של אורך המשפטים	.205	.026

$R^2 = .507$

סיכום

- נמצאו מספר מאפייני טקסט המנבאים קושי של טקסט המקובצים לשלוש קטגוריות:
  - מאפיינים למורכבות תחבירית
  - מאפיינים לרמה לקסיקלית
  - מאפיינים נוספים
- מאפייני הקריאות שנמצאו במחקר זה אינם חופפים למאפיינים המשמשים בנוסחות קריאות קיימות (בשפה האנגלית).
- מודל ניבוי המשלב 5 מהמאפיינים מסביר 50% מהשונות ברמת הקושי של הטקסטים.

מגבלות המחקר

- מדגם קטן יחסית
- שונות קטנה מאוד בין הטקסטים

מחקר עתידי

- מחקר בהיקף גדול יותר
- ניתוח בחתך של תחומי הדעת של הטקסטים
- זיהוי מדדי קריאות להערכת רמת הקושי של טקסטים מסוגים שונים
- בדיקת הגדרות נוספות של משתנה הקריטריון

תודות

צוות "פרויקט השפה העברית"

יונתן סער

ספי פומפיאן

עובדי התחום המילולי

מחלקת גרפיקה