המבנה הגורמי של שפה עברית כתובה

יעל שפרן, ענת בר-סימן-טוב

פרויקט השפה העברית (HLP), מאל"ו
הכנס השביעי של אפ"י, 2011, ירושלים

מה ההבדל בין שני הטקסטים?

ציון גבוה

.1

...לסיכום, אני מאמינה שבעניין הזיופים יש להתחשב באינטרסים של החברות הבינלאומיות אשר הן הנפגעות הישירות מתרבות זו. יש לצמצם את האפשרות לזיופים ע"י דרכים שונות כמו הגשת תביעות נגד המזייפים או להוריד באופן ניכר את מחירי המוצרים המקוריים. בשיטה זו, יהנו גם היצרנים וגם הצרכנים מן המוצרים בצורה היעילה ביותר.

ציון נמוך

.2

...במאמר זה יש הבטים שונים בנושא תופעת הזיופים בארץ מצד אחד תופעת הזיופים טובה כי זה יותר זול להוריד שירים מהאינטרנט מאשר לקנות דיסק מכורי וזה הרבה יותר זול לקנות ג'ינס מזויף של דיזל מאשר לקנות את המקורי ב- 900 ₪ !!! וזה גם יותר זול לקנות סרט צרוב מאשר מקורי

* מתוך קורפוס חיבורים שכתבו תלמידי י"ב דוברי עברית

מטקסט לאפיון כמותי

טקסט

מאפייני טקסט כמותיים

	% מילים נדירות	אורך ממוצע של משפט	% מילים באורך 6 ומעלה	% פעלים בסביל	ממוצע מילות קישור במשפט	גיוון לקסמות
ציון גבוה	0.37	14.7	0.33	0.05	2.5	7.3
ציון נמוך	0.28	17.4	0.21	0	1.2	5.8

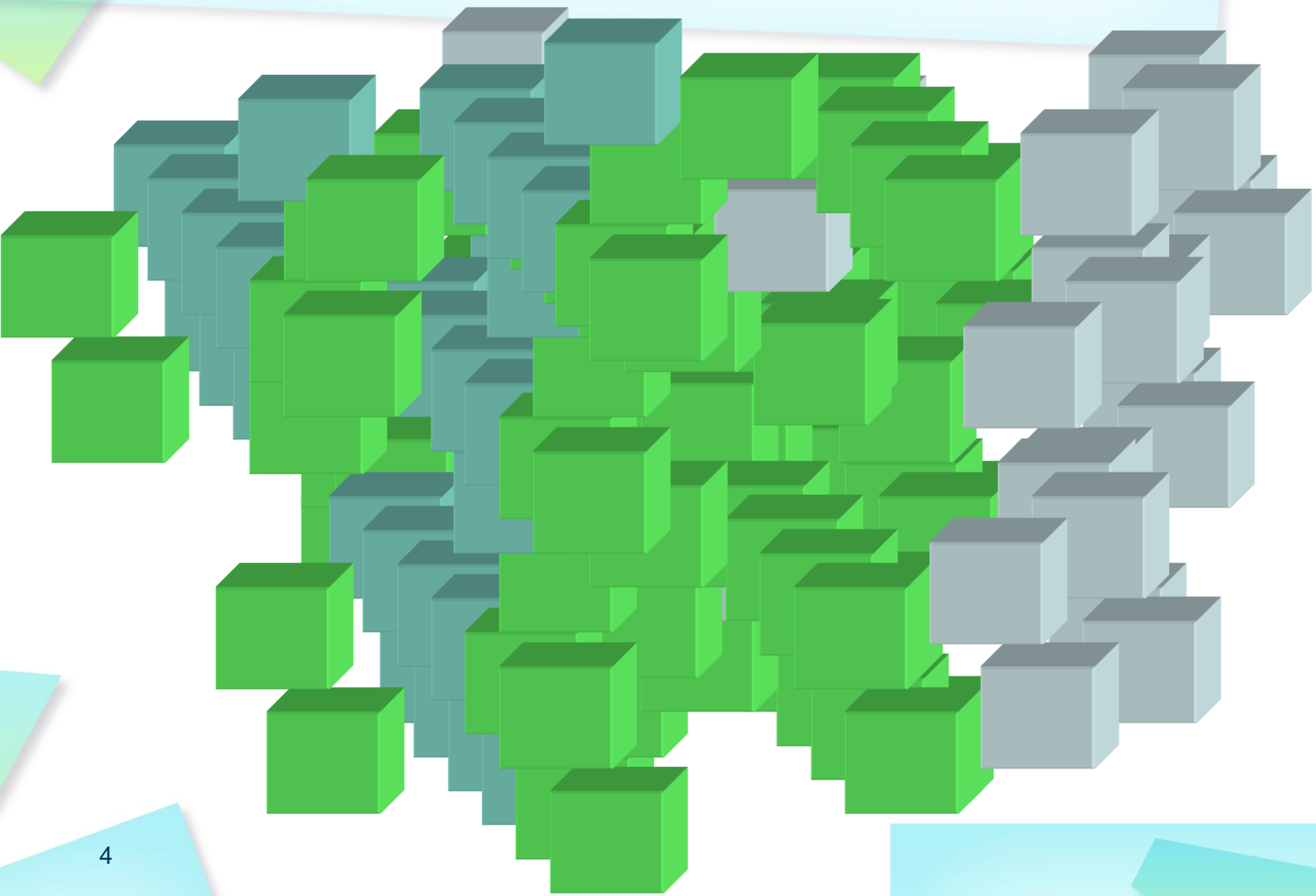
1. ציון גבוה

...לסיכום, אני מאמינה שבעניין הזיופים יש להתחשב באינטרסים של החברות הבינלאומיות אשר הן הנפגעות הישירות מתרבות זו. יש לצמצם את האפשרות לזיופים ע"י דרכים שונות כמו הגשת תביעות נגד המזייפים או להוריד באופן ניכר את מחירי המוצרים המקור
היצרני
בצורה

2. ציון נמוך

...במאמר זה יש הבטים שונים בנושא תופעת הזיופים בארץ מצד אחד תופעת הזיופים טובה כי זה יותר זול להוריד שירים מהאינטרנט מאשר לקנות דיסק מכורי וזה הרבה יותר זול לקנות ג'נס מזויף של דיזל מאשר לקנות את המקורי ב- 900 ₪ !!! וזה גם יותר זול לקנות סרט צרוב מאשר מקורי

קיימים מאות מדדים כמותיים שניתן להשתמש בהם לאפיון טקסט



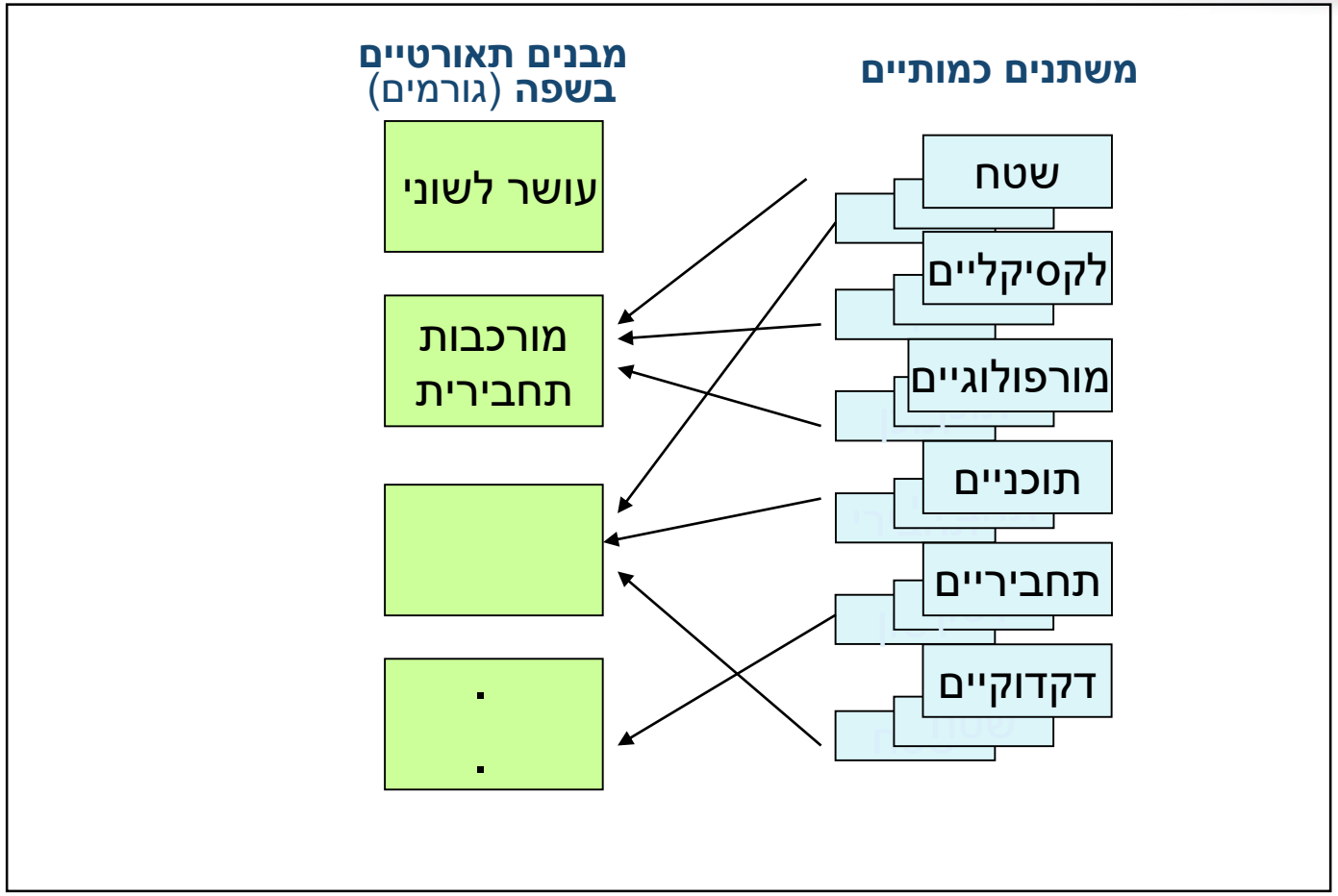
דוגמאות למאפיינים כמותיים של טקסט

משתני שטח*	לקסיקליים	מורפולוגיים	תוכניים	תחביריים	דקדוקיים
▪ # מילים ▪ ממוצע אורכי משפט ▪ ממוצע אורכי מילים	▪ % המילים הנדירות בטקסט ▪ ממוצע השכיחויות של מילים בטקסט	▪ % חלק דיבר מסוים (19) ▪ % בניין מסוים (8) ▪ % פעלים בסביל	▪ % מילים מקטגוריית תוכן נתונה	▪ # שעבודים ממוצע למשפט	▪ % שגיאות הכתיב ▪ % שגיאות התאם מין ומספר

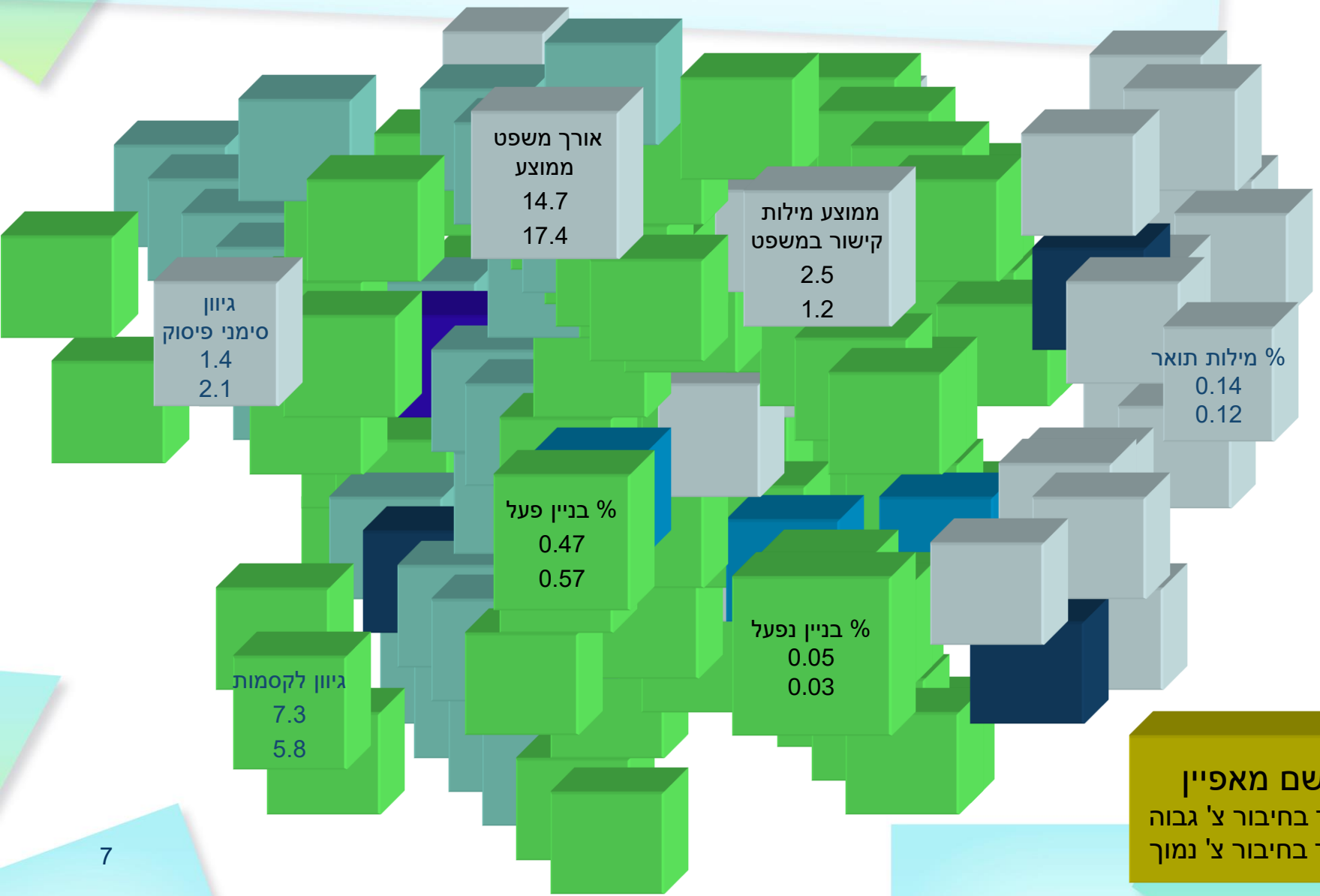
עד כה הגדרנו 133 מאפיינים

* משתני שטח - מאפיינים סטטיסטיים שאינם מצריכים ידע לשוני, ויכולים להעיד בעקיפין על ממדים לשוניים

כיצד עוברים מאוסף גדול של מאפיינים למבנה קוהרנטי ומצומצם?



התרשמות לא שיטתית מקשרים בין מאפיינים



בחינה שיטתית של קשרים בין מאפיינים:
ניתוח גורמים בשיטה לא אורתוגונאלית

- **המשתנים**: **133** מאפיינים כמותיים (לכל טקסט)
- **המדגם**: 3 קורפוסים של טקסטים

	קורפוס M1 טקסטים ערוכים ממקורות שונים	**חיבורים: תלמידי י"ב** דוברי עברית כשפת אם	**חיבורים: מבחן יע"ל** דוברי עברית כשפה שנייה
N	639	668	985

מהלך העבודה: ניתוח גורמים דו שלבי

- **שלב א'**
- ניתוח גורמים של **133** מאפיינים ב-3 קורפוסים
- חקירת ה"התנהגות" של כל מאפיין
  - אילו מאפיינים מתקבצים יחד על אותו גורם (הטעינות על הגורמים)
  - מה ההתפלגות של כל מאפיין
  - מה המתאם של כל מאפיין עם רמת קושי של טקסט או עם איכות כתיבה
- צמצום וטיוב המאפיינים
- **שלב ב'**
- ניתוח גורמים מחודש של **72** מאפיינים ב-3 קורפוסים
- הגדרה סופית של גורמים ותתי-גורמים

שלב א': ניתוח גורמים

משתנה (14 משתנים)	מתאמים עם ציון			הגורם עליו טעון המשתנה							המטען של המשתנה						
				M1			חיבורי י"ב			יע"ל	M1			חיבורי י"ב			יע"ל
	M1	חיבורי י"ב	חיבורי יע"ל	מלא	מדגם 1	מדגם 2	מלא	מדגם 1	מדגם 2	מלא	מלא	מדגם 1	מדגם 2	מלא	מדגם 1	מדגם 2	מלא
מספר מילים (תמניות)	0.69	0.64	0.66	2	2	2	2	2	2	3	1	1	1	1	1	1	1
מספר מחרוזות (מילים, מספרים וסימני פיסוק)	0.69	0.64	0.68	2	2	2	2	2	2	3	0.9	1	0.9	1	0.99	0.99	0.99
מספר מילים שונות (תבניות)	0.71	0.67	0.72	2	2	2	2	2	2	3	1	1	1	0.9	0.91	0.92	0.92
מספר משפטים	0.60	0.56	0.35	2	2	2	2	2	2	4	0.7	0.8	0.7	0.7	0.76	0.71	-0.37
יחס תבניות לתמניות - מחרוזות	-0.45	-0.39	-0.10	2	2	2	2	2	2	3	-0.4	-0.3	-0.4	-0.8	-0.6	-0.7	-0.33
יחס תבניות לתמניות - לקסמות	-0.49	-0.48	-0.27	2	2	2	2	2	2	3	-0.4	-0.5	-0.5	-0.9	-0.7	-0.8	-0.55
גיוון מחרוזות	0.79	0.73	0.74	2	2	2	2	2	2	3	0.9	0.9	0.9	0.7	0.8	0.74	0.88
גיוון לקסמות	0.79	0.72	0.74	2	2	2	2	2	2	3	0.9	0.9	0.8	0.6	0.66	0.61	0.8

שלב א' 1: ניפוי מאפיינים

קריטריונים לניפוי מאפיין:

- דמיון בדפוס המשקולות
- דמיון בדפוס המתאמים עם הציון
- עקביות במדדים בתוך קורפוסים ומעבר לקורפוסים
- ייצוג של ישות לשונית דומה

משתנה (14 משתנים)	מתאמים עם ציון			הגורם עליו טעון המשתנה							המטען של המשתנה						
				M1			חיבורי י"ב			יע"ל	M1			חיבורי י"ב			יע"ל
	M1	חיבורי י"ב	חיבורי יע"ל	מלא	מדגם 1	מדגם 2	מלא	מדגם 1	מדגם 2	מלא	מלא	מדגם 1	מדגם 2	מלא	מדגם 1	מדגם 2	מלא
מספר מילים (תמניות)	0.69	0.64	0.66	2	2	2	2	2	2	3	1	1	1	1	1	1	1
מספר מחרוזות (מילים, מספרים וסימני פיסוק)	0.69	0.64	0.68	2	2	2	2	2	2	3	0.9	1	0.9	1	0.99	0.99	0.99
מספר מילים שונות (תבניות)	0.71	0.67	0.72	2	2	2	2	2	2	3	1	1	1	0.9	0.91	0.92	0.92
מספר משפטים	0.60	0.56	0.35	2	2	2	2	2	2	4	0.7	0.8	0.7	0.7	0.76	0.71	-0.37
יחס תבניות לתמניות - מחרוזות	-0.45	-0.39	-0.10	2	2	2	2	2	2	3	-0.4	-0.3	-0.4	-0.8	-0.6	-0.7	-0.33
יחס תבניות לתמניות - לקסמות	-0.49	-0.48	-0.27	2	2	2	2	2	2	3	-0.4	-0.5	-0.5	-0.9	-0.7	-0.8	-0.55
גיוון מחרוזות	0.79	0.73	0.74	2	2	2	2	2	2	3	0.9	0.9	0.9	0.7	0.8	0.74	0.88
גיוון לקסמות	0.79	0.72	0.74	2	2	2	2	2	2	3	0.9	0.9	0.8	0.6	0.66	0.61	0.8

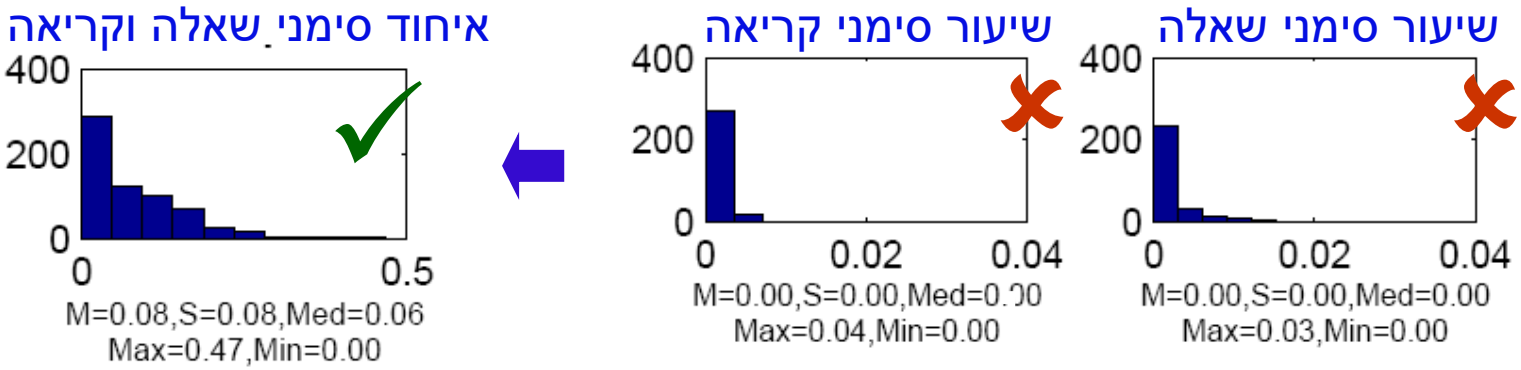
14 מאפיינים

⬇

8 מאפיינים

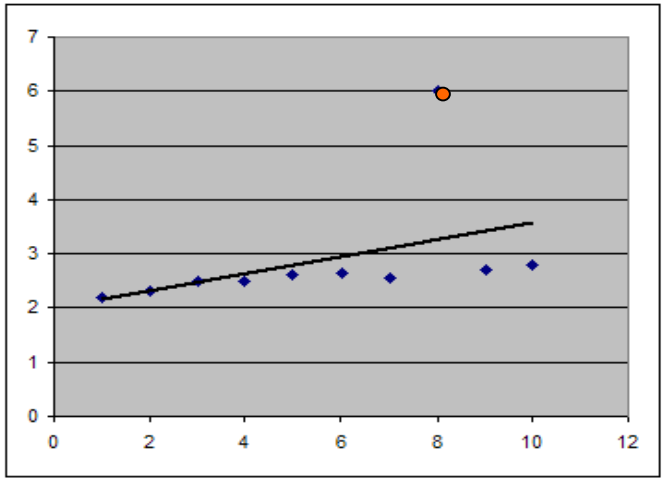
שלב א' 2: בדיקת שונות המאפיינים
איחוד וניפוי

- מאפיינים בעלי שונות נמוכה אוחדו עם מאפיינים דומים מבחינת תפקיד לשוני. למשל:
  - אוחדו סימני פיסוק ייחודיים בסוף משפט (!+?)
  - אוחדו סימני פיסוק באמצע משפט (, + ; + :).

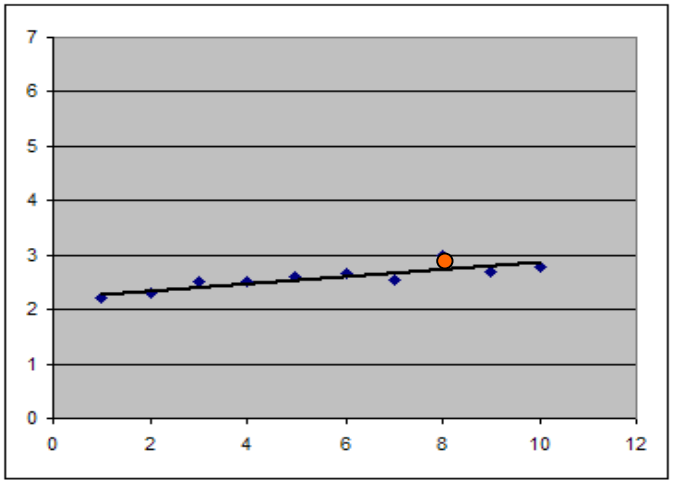


שלב א' 3: טיפול בערכים חריגים

קיים ערך חריג
R=0.42



תיקון לערך חריג
R=0.86



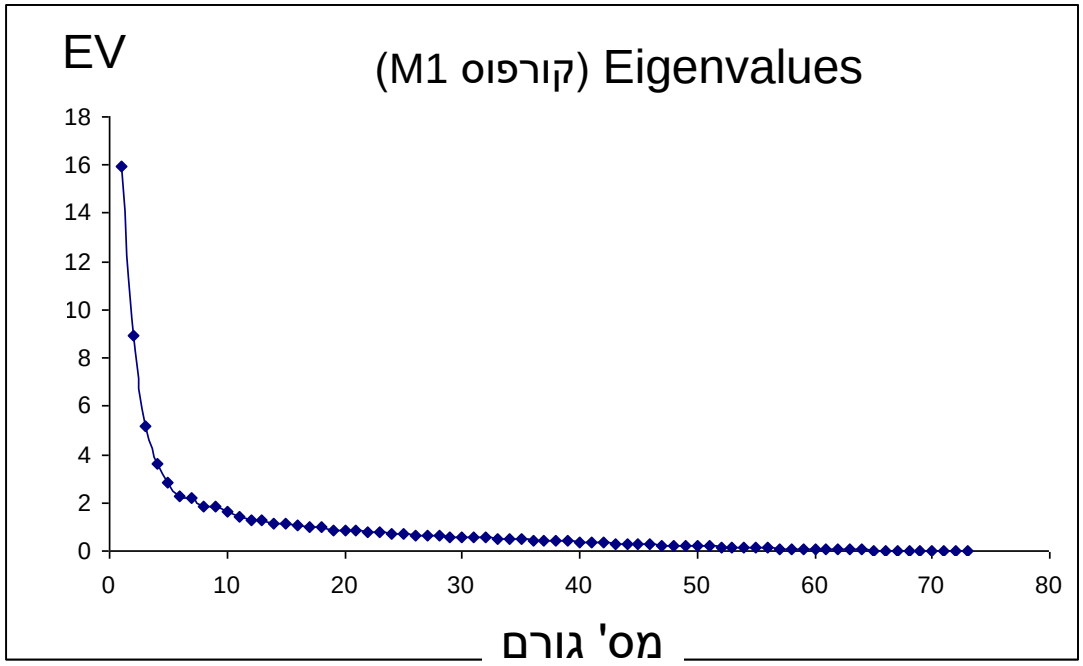
- מקורות לערכים חריגים: שונות נמוכה של המאפיין או טקסט קצר מאוד.
- <u>פתרון</u>: הגבלה של טווח ההשתנות של ציוני התקן של כל מאפייני הטקסט **ל- 3 ± סטיות תקן**

סיכום שלב א'

- הוסרו 73 מאפיינים מתוך 133
- נוספו 12 משתנים (שופצו או אוחדו)
- סה"כ התקבלו **72** משתנים

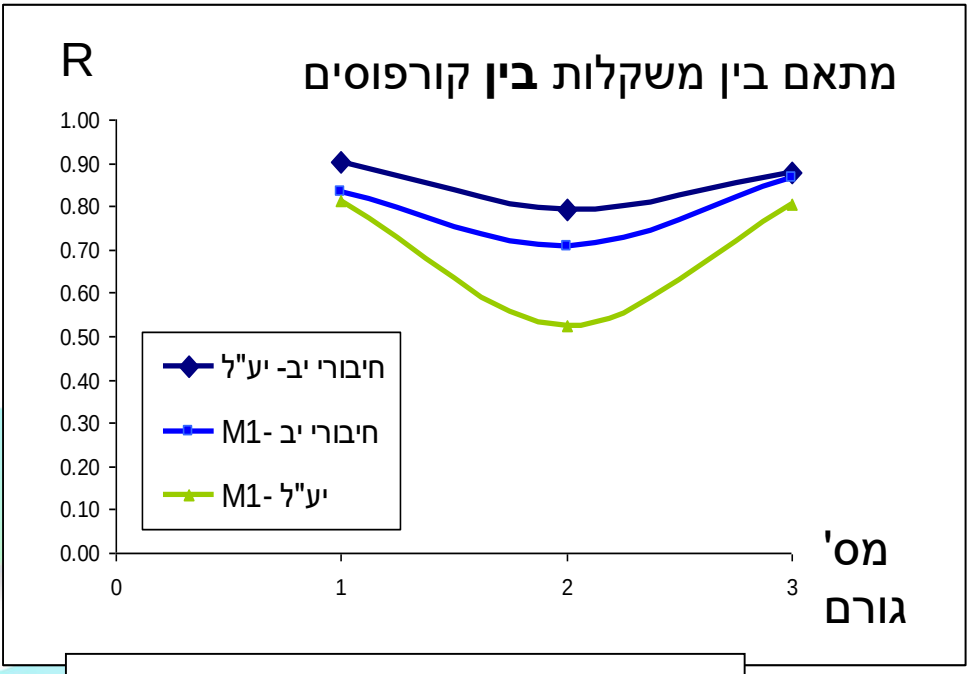
שלב ב': ניתוח גורמים של 72 מאפיינים

- **כמה גורמים משמעותיים קיימים?**
- 15 גורמים מסבירים 72%-61% מהשונות בשלושת הקורפוסים

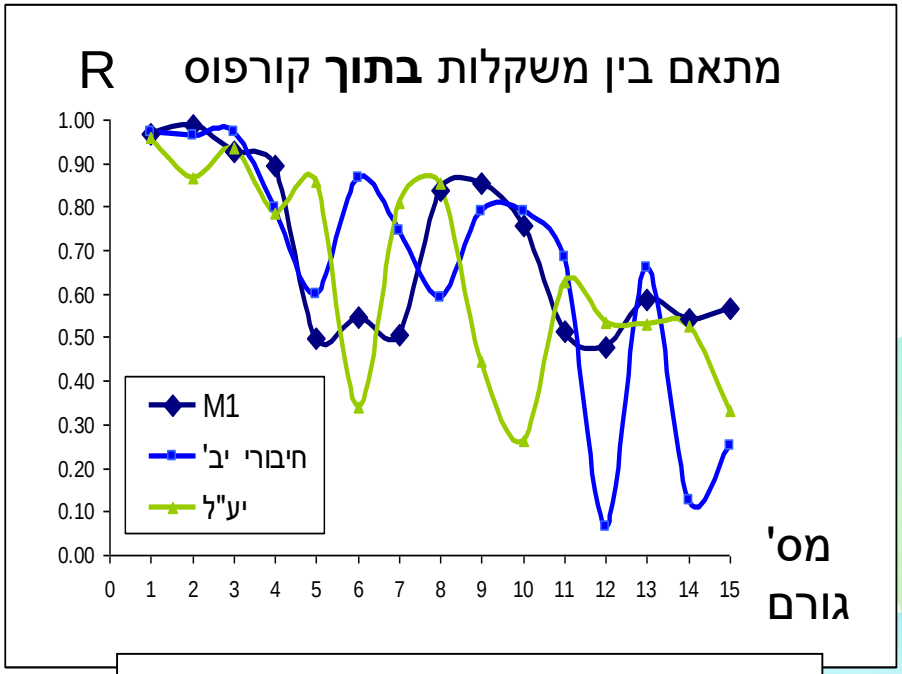


שלב ב': ניתוח גורמים של 72 מאפיינים

- **אילו גורמים הם יציבים (בתוך כל קורפוס ובין קורפוסים)?**
- תיקוף צולב של הגורמים (בתוך כל קורפוס ובין קורפוסים)
- תוצאות: 4-3 הגורמים הראשונים יציבים ועקביים



3 גורמים ראשונים (r= 0.52 - 0.9)



4 גורמים ראשונים (r= 0.79 - 0.97)

גורם 1: שכיחות מילים ודחיסות תוכנית

גורם	תת-גורם	משתנה
1. שיעור מילים שכיחות ודחיסות תוכנית	**1.1 שיעור מילים שכיחות**	שכיחות ממוצעת של מחרוזת
		שכיחות ממוצעת של לקסמה
		אורך ממוצע של מחרוזת
		שיעור מחרוזות באורך 10 ומעלה
		שיעור מחרוזות באורך 6 ומעלה
	1.2 דחיסות תוכנית	שיעור מילות פונקציה
		שיעור מילות תוכן
	1.3 שמות תואר	שיעור שמות תואר
	1.4 מילים עם תחיליות	שיעור מילים עם תחילית
	1.5 כינוי רומז	שיעור כינויים רומזים
	1.6 כינוי גוף	שיעור כינויי גוף
	1.7 כמתים	שיעור כמתים
	1.8 ז'אנר ספרותי לילדים - פועל בתחילת משפט	שיעור משפטים הפותחים בפועל
		שיעור לקסמות ממוצע למחרוזת

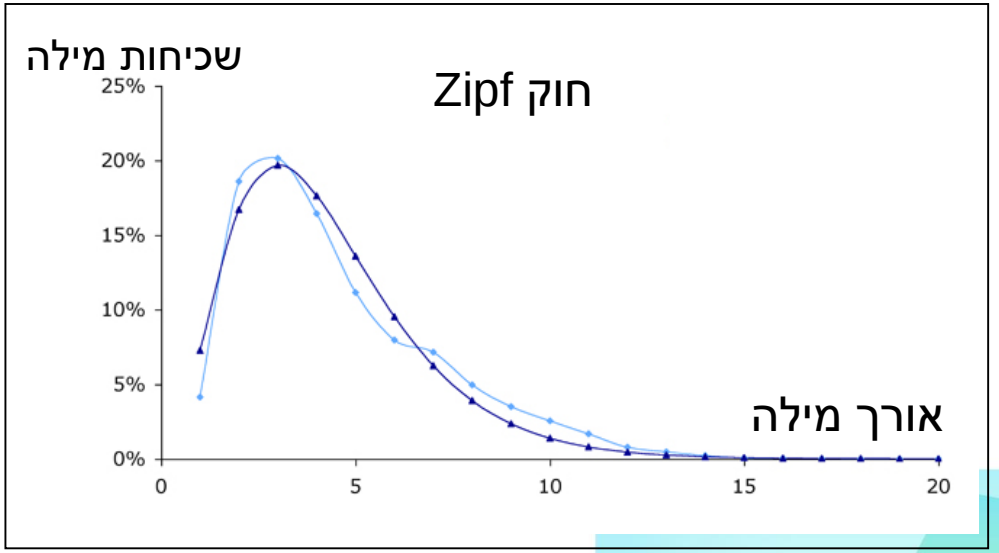
גורם 1: הקשר בין שכיחות מילים לאורך מילים

תת-גורם	משתנה	משקל
1.1 שיעור מילים שכיחות	שכיחות ממוצעת של מחרוזת	0.87
	שכיחות ממוצעת של לקסמה	0.91
	אורך ממוצע של מחרוזת	-0.89
	שיעור מחרוזות באורך 6 ומעלה	-0.8

מאפיינים לקסיקליים

מאפייני שטח

תת הגורם תואם את חוק Zipf: ככל שמילה קצרה יותר כך היא שכיחה יותר



גורם 2: כמות הטקסט וגיוונו

גורם	תת-גורם	משתנה	
2. כמות הטקסט וגיוונו	2.1 כמות מחרוזת	לוגריתם של מספר המילים	S_LETTER_STRING_LOG_CNT
		מספר מילים שונות	S_LETTER_STRING_TYPES_CNT
	2.2 גיוון לקסיקלי	גיוון מחרוזות	S_STRING_DIVIRSITY
		גיוון לקסמות	M_LEXEME_DIVIRSITY
	2.3 הכרת מילים נדירות שונות	שיעור מחרוזות נדירות (תבנית)	S_TYPE_FREQ_AVG
		שיעור לקסמות נדירות (תבנית)	S_TYPE_FREQ_LOW
		שכיחות ממוצעת של מחרוזות (תבנית)	S_STRING_TYPE_FREQ_AVG
		שכיחות ממוצעת של לקסמה (תבנית)	M_LEXEME_TYPE_FREQ_AVG
	2.4 גיוון לשוני מדד Z	גיוון לשוני מדד Z	S_FREQ_CURVE_ZIPF
	2.5 גיוון לשוני מדד D	גיוון לשוני מדד D	S_TTR_CURVE_D
	2.6 גיוון מילות יחס	גיוון מילות יחס	M_PREPOSITION_DIVIRSITY

גורם 3: משפטים ארוכים / מורכבות תחבירית בטקסטים תקניים

גורם	תת-גורם	משתנה
3. משפטים ארוכים / מורכבות תחבירית בטקסטים תקניים	**3.1** משפטים ארוכים	אורך משפט ממוצע
		שיעור המשפטים הארוכים במיוחד
		שיעור המשפטים הקצרים במיוחד
		סטיית התקן של אורכי המשפטים
	3.2 מורכבות תחבירית בטקסטים תקניים	ממוצע מילות חיבור ושעבוד במשפט
		ממוצע מילות יחס במשפט
	3.3 מספר שלילות גבוה במשפט	שיעור המשפטים בעלי מספר שלילות גדול מ-1

גורם 3: הבדל בין טקסט ערוך לחיבור

מתאמים בין ציון חיבור / גיל קורא של טקסט למאפיינים

3.1 משפטים ארוכים | 3.2 מורכבות תחבירית

מתאם עם ציון	אורך משפט ממוצע	שיעור המשפטים הארוכים במיוחד	ממוצע מילות חיבור ושעבוד במשפט	ממוצע מילות יחס במשפט	גיוון סימני פיסוק
טקסט ערוך נרטיבי M1	0.40	0.46	0.32	0.40	-0.22
חיבורי י"ב	-0.04	-0.25	-0.05	0.02	0.17

משפטים ארוכים

חיבורים: חוסר פיסוק

טקסט ערוך: משפט מורכב

ולסיום... מעבר לתת גורמים

טקסט

מאפייני טקסט כמותיים: תת גורמים

	1.1 שיעור מילים שכיחות	2.2 גיוון לקסיקלי	3.2 מורכבות תחבירית
ציון גבוה	- 1.17	2.3	- 0.7
ציון נמוך	0.2	- 3.1	- 1.83

1. ציון גבוה

...לסיכום, אני מאמינה שבעניין הזיופים יש להתחשב באינטרסים של החברות הבינלאומיות אשר הן הנפגעות הישירות מתרבות זו. יש לצמצם את האפשרות לזיופים ע"י דרכים שונות כמו הגשת תביעות נגד המזייפים או להוריד באופן ניכר את מחירי המוצרים המקורי היצרנ בצורה

2. ציון נמוך

...במאמר זה יש הבטים שונים בנושא תופעת הזיופים בארץ מצד אחד תופעת הזיופים טובה כי זה יותר זול להוריד שירים מהאינטרנט מאשר לקנות דיסק מכורי וזה הרבה יותר זול לקנות ג'נס מזויף של דיזל מאשר לקנות את המקורי ב- 900 ₪ !!! וזה גם יותר זול לקנות סרט צרוב מאשר מקורי

סיכום
המבנה הגורמי של שפה עברית כתובה
או
איך מאפיינים טקסט?

- ניתוח גורמים - נושאים מתודולוגיים
  - ניפוי מאפיינים
  - בדיקת התפלגות המאפיינים ואיחוד מאפיינים
  - טיפול במקרים חריגים
- גורמים בשפה העברית:
  - סה"כ 15 גורמים, שחולקו ל: 12 תת גורמים + 26 מאפיינים בדידים.
  - שלושת הגורמים הראשונים הם היציבים ביותר

תת גורם	שם הגורם
1.1	שיעור מילים שכיחות
1.2	דחיסות תוכנית
2.1	כמות מחרוזת
2.2	גיוון לקסיקלי
2.3	הכרת מילים נדירות שונות
3.1	משפטים ארוכים
3.2	מורכבות תחבירית בטקסטים תקניים

תודה