סימפוזיון: ניתוח אוטומטי של טקסטים בעברית – יישומים במחקר ובהערכה

פרויקט השפה העברית

Hebrew Language Project (HLP)

מאל"ו

הכנס השביעי של אפי, 2011, ירושלים

סימפוזיון: ניתוח אוטומטי של טקסטים בעברית – יישומים במחקר ובהערכה

- מבוא-כלים ושיטות לניתוח ממוחשב של טקסטים בשפה העברית - יואב כהן
- המבנה הגורמי של מאפייני טקסט בשפה העברית הכתובה - יעל שפרן, ענת בר-סימן-טוב
- התרומה הדיפרנציאלית של מאפייני טקסט להערכה ממוחשבת של חיבורים מסוגים שונים - ענת בן-סימון
- הערכת רמות קושי של טקסטים המשמשים בהערכת יכולת הבנת הנקרא-באמצעות מאפיינים סטטיסטיים ומורפולוגיים של הטקסט - דנה רובינשטיין, יעל שפרן, ענת בן-סימון

מבוא - כלים ושיטות לניתוח ממוחשב של טקסטים בשפה העברית

יואב כהן

- פרויקט השפה העברית (HLP), מאל"ו

הכנס השביעי של אפי, 2011, ירושלים

הקדמה

- המטרה העיקרית של הפרויקט:

ניצול כלים של בלשנות חישובית לצורך הערכה של תוצרי כתיבה.

- משתתפים ותורמים לפרויקט:
  - ענת בן-סימון – מנהלת הפרויקט
  - יעל שפרן, ענת בר-סימן-טוב, גלי נוטי, אפי
  - מירה חובב, אפרת מילר, מרק שובמן, דוד קשתן, אלעד דינור, אייל שוורץ, רועי רייכרט

- בארה"ב קיימים מספר פרויקטים:
  - **Page-PEG** – 1966, 1994, 1995
  - **eRater** – 1997 ETS, וכן Burstein et al. 1998,1999,2000,2001
  - **Intellimetric** – 1997, 2001 Elliot
  - **IEA** – 1997 Landauer et al.
- חלק מן הכלים אינם תלויי שפה (Cohen, Ben-Simon & Hovav, 2003)

הקדמה

ייחודיות של השפה העברית:

- שיטות כתיב שונות
  - כתיב מנוקד
  - "כתיב חסר"
  - ניקוד חלקי
  - כתיב חסר ניקוד ("כתיב מלא")
- הטיות לסוגיהן
  - הטיות השורשים ליצירת פעלים
  - כינויים חבורים
- מוספיות
  - תחיליות – אותיות שימוש, ה הידיעה, ו החיבור
  - סופיות – ה המגמה

ריבוי מחרוזות, ריבוי משמעויות

- בעברית כמה **מיליוני** מילים
  - שורשים:
    - לכל שורש: 2עד 7 בניינים 3 זמנים 3 גופים 2 צורות ריבוי 2 מינים --> 72 עד 252 צורות
    - וכן: כינויים חבורים: כינוי המושא (4 לכל הטייה) --> 280 - 1000 צורות
    - וכן תחיליות: ו ההיפוך, ש, כש -->840 - 3000
    - בעברית מקראית 2000 שורשים, שפת המשנה הוסיפה 800
    - מספר הטיות השורשים המינימלי הוא איפוא: 2.3 מיליון!
  - שמות
    - 24 כינויי קניין (2 צורות ריבוי, 12 כינויי שייכות)
    - 4 עד 5 תחיליות (ו,ל,ש,מ,ב...)
    - אם יש 30,000 שמות -->3,000,000 צורות שונות של שמות

ריבוי הומוגרפים

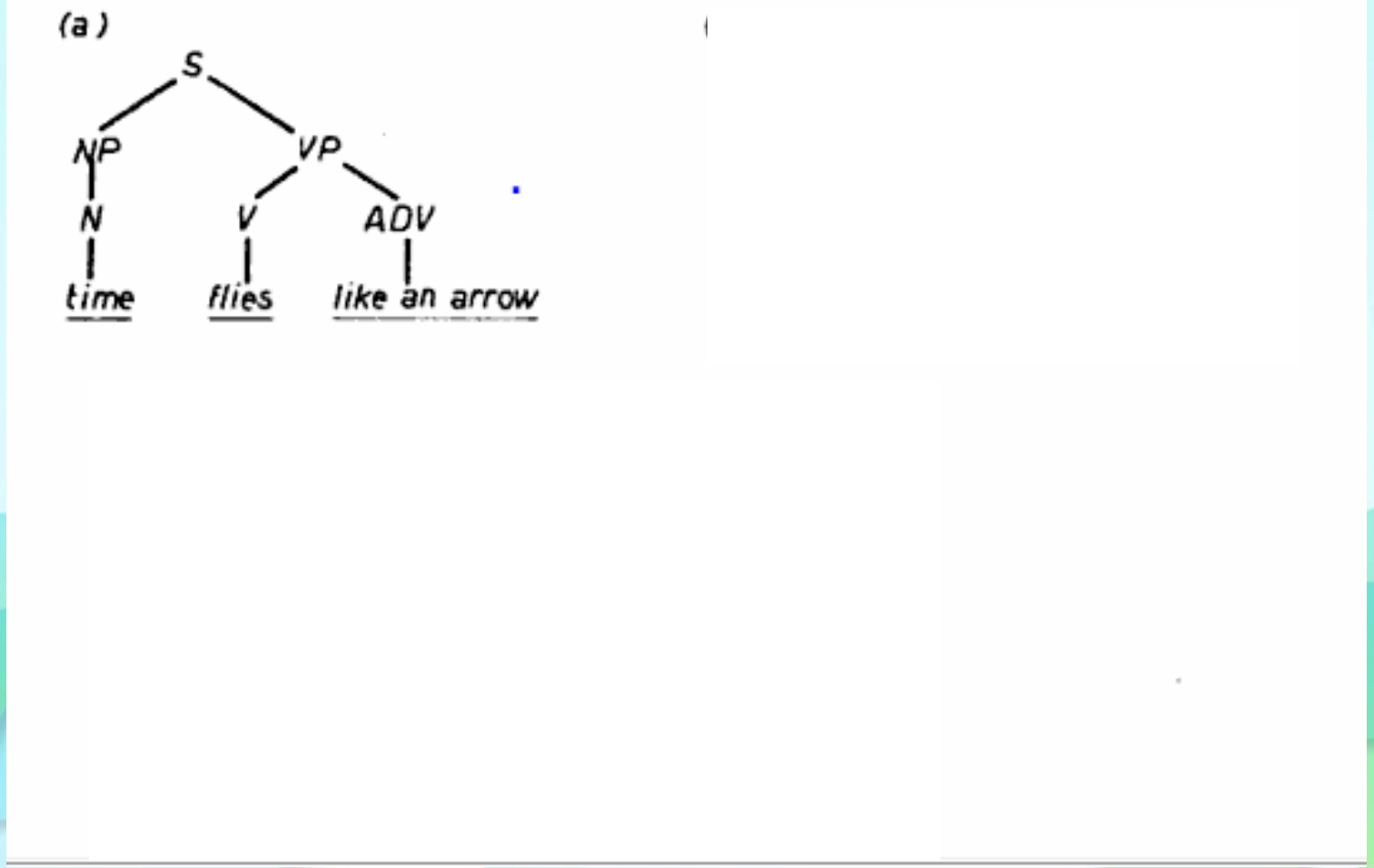
ספר

- SEFER
- SAFAR
- SAPAR
- SFAR
- SAPER
- (SUPAR)
- (SIPPER)

עמימות

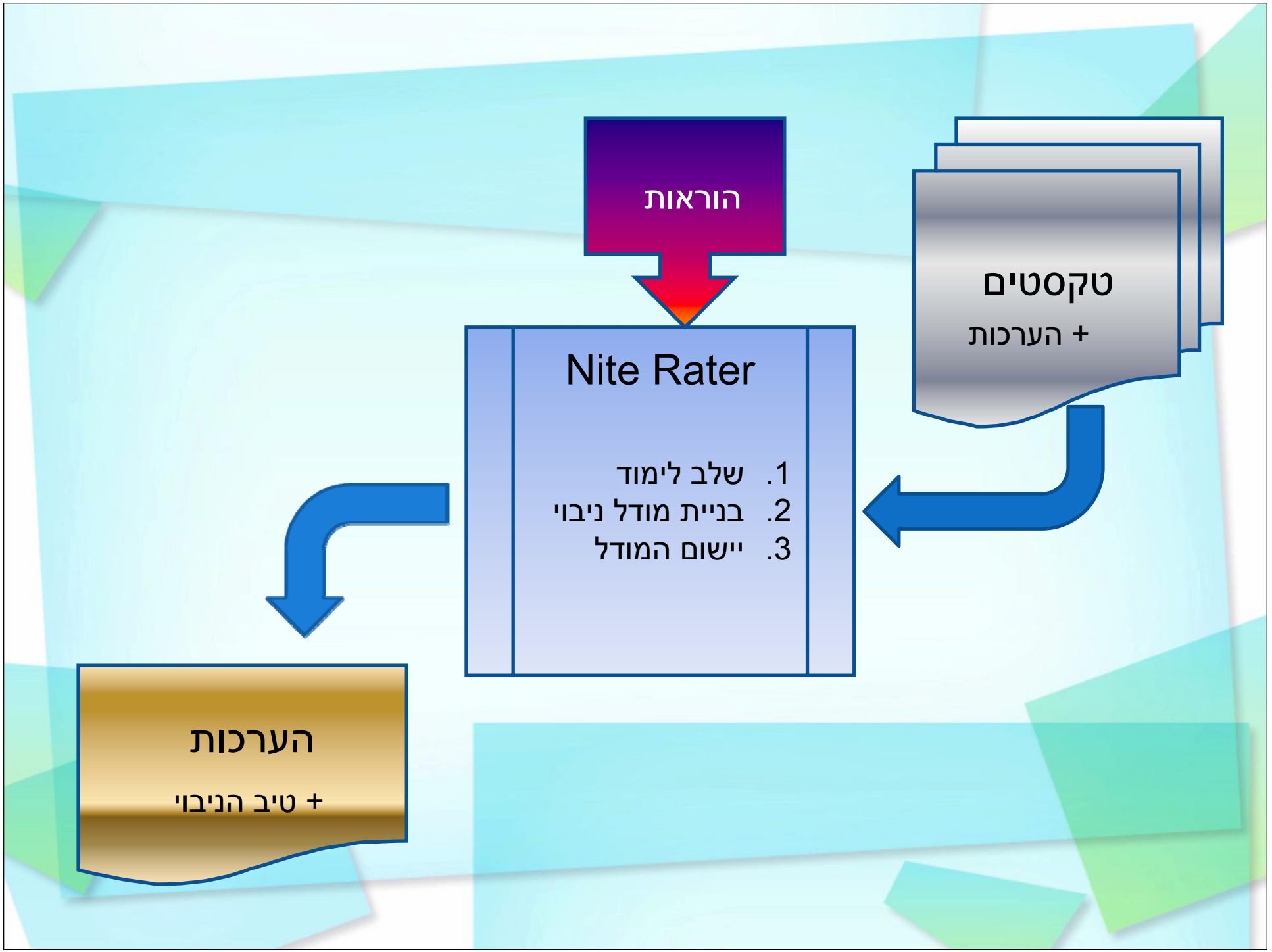
"Time flies like an arrow"

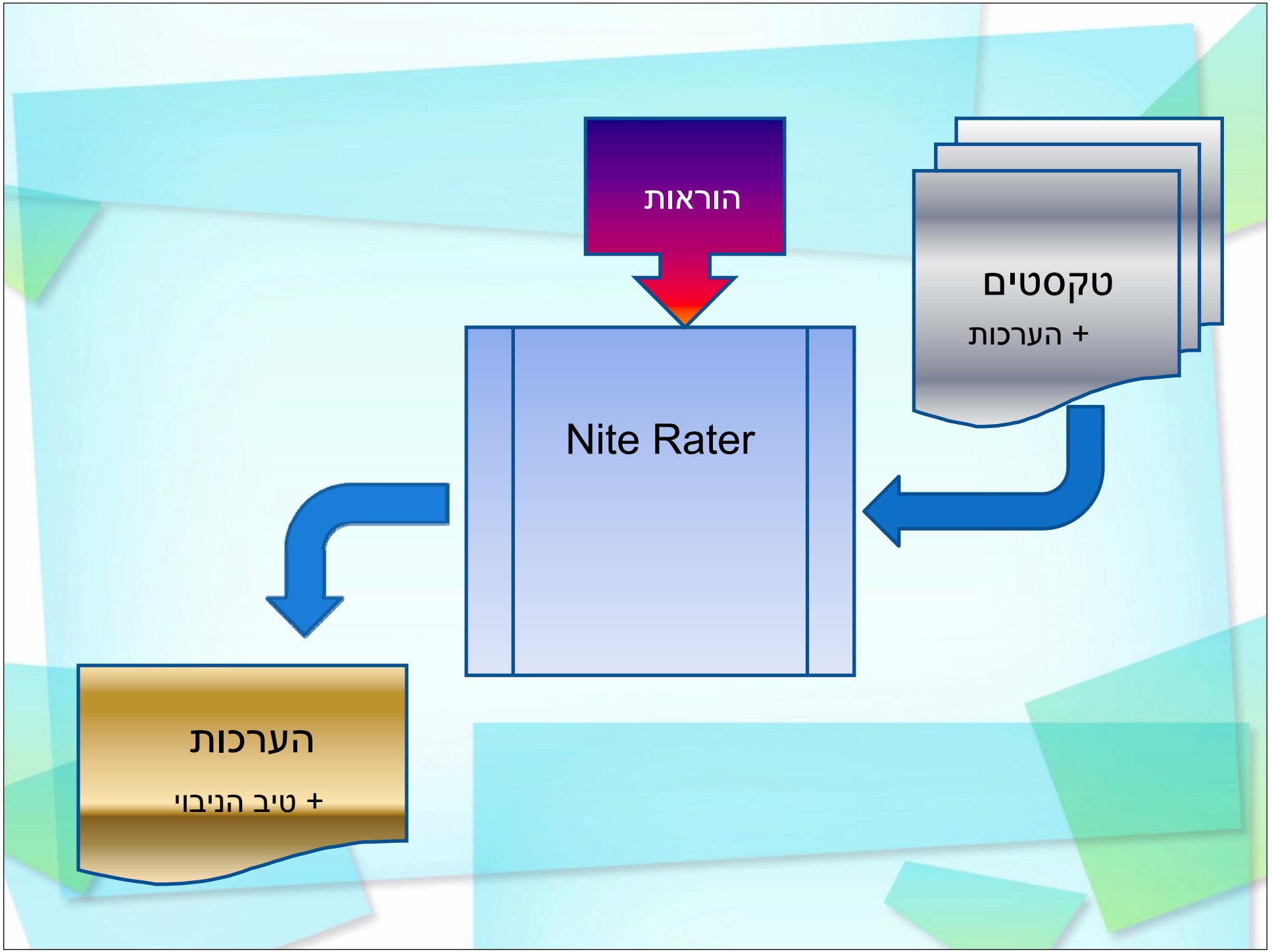
(Oetinger, 1966)



ריבוי צורות הקריאה

"כשהתחלתי לחשוב על ייסודו של כתב עת זה, בדקתי ומצאתי שלמעלה מ-1,100 אישה ואיש בארץ עוסקים בפסיכואנליזה במסגרות, בחברות, ובמכונים שונים" (גבי שפלר, 2010)

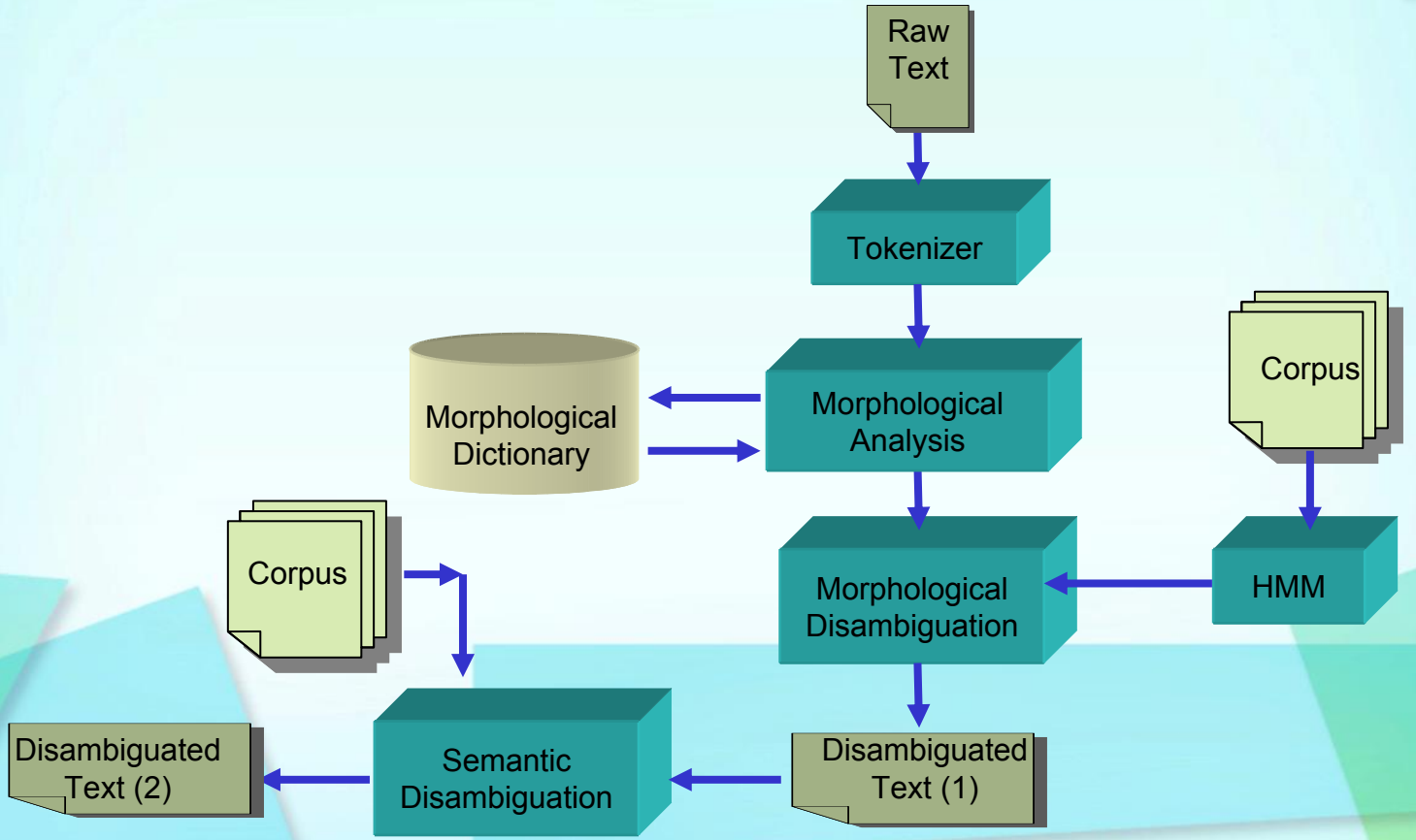

הוראות
טקסטים
+ הערכות
Nite Rater
1. שלב לימוד
2. בניית מודל ניבוי
3. יישום המודל
הערכות
+ טיב הניבוי


הוראות
טקסטים
+ הערכות
Nite Rater
הערכות
+ טיב הניבוי

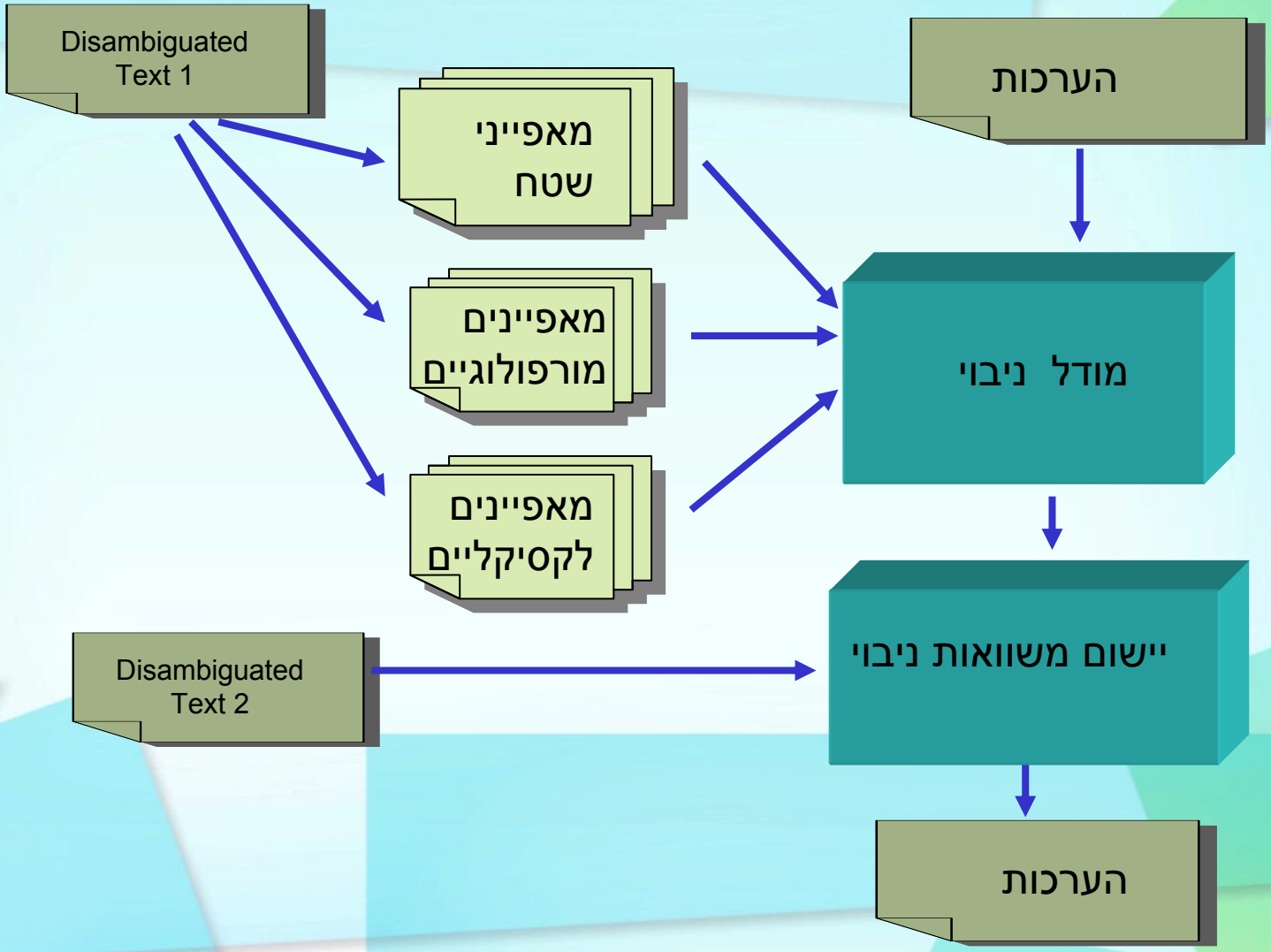
Nite Rater

- שלב לימוד -- בכלים של בלשנות חישובית
  - הפרדת הטקסט ליחידות -- Tokenizer
  - איפיון הטקסט -- תוכנת איפיון טקסטים: CAI, איפיונים "פיזיקליים"/"סטטיסטיים"/"מאפייני שטח"
  - פיענוח הטקסט - מנתח מורפולוגי
    - הפגת עמימות – מפיג עמימות אוטומטי
    - הפגת עמימות סמנטית
  - איפיון הטקסט: איפיונים מורפולוגיים ולקסיקליים
    - שימוש בבניינים, כינויים חבורים וכד'
    - עושר הביטוי, משלב לשוני (שכיחות מילים)
- שלב הניבוי
  - משוואות רגרסיה
  - תיקוף צולב
  - יישום משוואות הרגרסיה

תיאור סכמטי של מפיג העמימות המורפולוגי והסמנטי



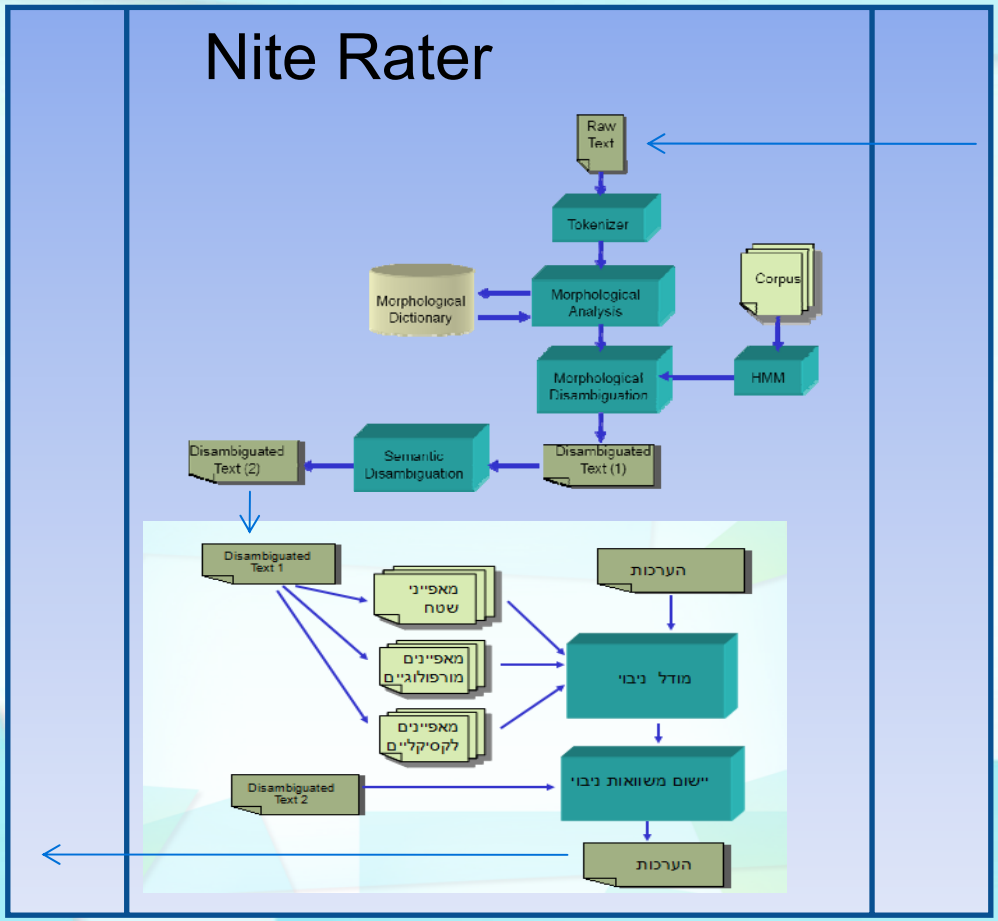
תיאור סכמטי של שלב הניבוי



הוראות

טקסטים
+ הערכות

Nite Rater



הערכות
+ טיב ניבוי

סיכום – "ארגז הכלים"

- Tokenizer
- מפיג עמימות מורפולוגית
- מילון מורפולוגי
- קורפוס מתויג
- מפיג עמימות סמנטית
- לקסמות
- איפיוני שטח
- איפיונים מורפולוגיים ולקסיקליים

עד כאן ההקדמה,

ועכשיו: מה לומדים מזה ומה עושים עם זה?

תודה רבה

Tokenizer

- תוכנה המפרידה בין מחרוזות, מספרים וסימני פיסוק.

לדוגמה:

"הספר עולה סה"כ 25 NIS."

	The book	costs	In total	25	NIS	.
	הספר	עולה	סה"כ	25	NIS	.
Type	Word	word	word	Number	word	punctuation
Sub-type	Hebrew	Hebrew	Acronym	Number	Foreign	Sentence ender

מאפיינים סטטיסטיים ב'מאפיין הטקסטים' (CAI)

קטגוריה	מספר	שם משתנה	תיאור
תו	1.1	**# תווים**	מספר התווים הכולל (כל סימני הדפוס ללא רווחים, סוף שורה וסוף פסקה)
	1.2	**# אותיות ומספרים**	מספר התווים המופיעים במילים
	1.3	**# מספרים**	מספר המספרים (לדוגמה, המספר 203 יספר כמספר אחד).
	1.4	**יחס מספרים למחרוזות**	היחס של מס׳ המספרים למס׳ המחרוזות
	1.5	**אורך ממוצע של מספר**	אורך ממוצע של מספר (למשל, אורכו של המספר 203, הוא 3)
	1.6	**# סימני פיסוק**	מס׳ סימני הפיסוק הכולל בטקסט
	1.7	**# סוגי סימני פיסוק**	מס׳ סימני הפיסוק השונים בטקסט
	1.8	**יחס סימני פיסוק למחרוזות**	שיעור המופע של כל אחד מסימני הדקדוק (כנ״ל פרופורציה : מספר מופעי הסי מכלל סימני הדקדוק בטקסט)
	1.9	**שיעור המופע היחסי של אותיות הא"ב**	שיעור המופע של כל אחת מ אותיות הא״ב בטקסט מתוך כלל האותיות
	1.10	**שיעור המופע היחסי של אותיות הא"ב שבראש מילה**	שיעור המופע של כל אחת מ אותיות הא״ב בראש מילה
	1.11	**שיעור המופע היחסי של אותיות הא"ב בסוף מילה**	שיעור המופע של כל אחת מאותיות הא״ב בסוף מילה

מאפיינים סטטיסטיים ב'מאפיין הטקסטים' (CAI)

קטגוריה	מספר	שם משתנה	תיאור
משפט	3.1	**# המשפטים**	מס' המשפטים הכולל בטקסט
	3.2	**אורך ממוצע של משפט (במחרוזות)**	אורך ממוצע של משפט במחרוזות
	3.3	**אורך ממוצע של משפט (בתווים)**	אורך ממוצע של משפט בתווים
	3.4	**ס.ת. של אורך משפט (במחרוזות)**	סטיית התקן של אורכי המשפטים (במילים).
	3.5	**ס.ת. של אורך משפט (בתווים)**	סטיית התקן של אורכי המשפטים (בתווים)
	3.6	**אורך המשפט הראשון..שישי בטקסט**	אורך של המשפט הראשון – השישי בטקסט

Morphological Analyzer

- תוכנה המפיקה עבור כל מחרוזת בטקסט את כל הניתוחים המורפולוגיים האפשריים המתאימים לכל הפירושים האפשריים של המחרוזת
- לדוגמה

יש לנתח את המשפט:

"מחקר חדש: נתגלה הורמון הגורם..."

פלט של המנתח המורפולוגי

"מחקר חדש: נתגלה הורמון הגורם..."

מס שורה	מס מלה	צורה נטויה	צורת יסוד	שורש	חלק דיבר	בניין/משקל	זמן	גוף לפועל	מספר	מין	נפרד/נסמך	תחילית
1	1	מחקר	מחקר		שם עצם	81			יחיד	זכר	נפ/נס	
1	1	מחקר	מחקר	חקר	בינוני	פיעל	הווה	נייטראלי	יחיד	זכר	נפ/נס	
1	1	מחקר	חקר	חֶקֶר	שם עצם	292			יחיד	זכר	נפ/נס	מ
1	1	מחקר	מחו&קר	חקר	בינוני	פועל	הווה	נייטראלי	יחיד	זכר	נפ/נס	
1	1	מחקר	חקר	חקר	מקור	פיעל					נסמך	מ
1	1	מחקר	חקו&ר	חקר	מקור	פעל					נסמך	מ
1	2	חדש	חדש		תואר	662	אפיון		יחיד	זכר	נפ/נס	
1	2	חדש	חי&דש	חדש	פועל	פיעל	עבר	שלישי	יחיד	זכר		
1	2	חדש	חו&דש	ח ֹדֶש	שם עצם	297			יחיד	זכר	נפ/נס	
1	2	חדש	חדש	חדש	פועל	פיעל	ציווי	שני	יחיד	זכר		
1	2	חדש	חו&דש	חדש	פועל	פועל	עבר	שלישי	יחיד	זכר		
1	2	חדש	חדש	חדש	מקור	פיעל					נסמך	
1	3	:	נקודותיים		סימן פיסוק							
1	4	נתגלה	נתגלה	גלה	פועל	התפעל	עבר	שלישי	יחיד	זכר		
1	4	נתגלה	נתגלה	גלה	פועל	התפעל	עתיד	ראשון	רבים	נייטראלי		
1	5	הורמון	הורמון		שם עצם	3008			יחיד	זכר	נפרד	
1	6	הגורם	גורם	גוֹרֶם	שם עצם	174			יחיד	זכר	נפרד	ה
1	6	הגורם	גורם	גרמ	בינוני	פעל	הווה	נייטראלי	יחיד	זכר	נפרד	ה
1	6	הגורם	גורם	גור	מקור	פעל					נפרד	ה
1	6	הגורם	גו&רם	גרר	מקור	פעל					נפרד	ה

מה וכיצד יודע המנתח המורפולוגי?

המנתח המורפולוגי של NITE

מספר חלקי הדיבר העיקריים

- **פעלים**: 3,295 שורשים (369,933 מוטים)
  - **שמות עצם**: 15,341
  - **שמות תואר**: 2,489
  - **תוארי פועל**: 302
  - **שמות פרטיים**: 6,488
  - **מילות יחס**: 123
  - **סך הכול מחרוזות**: 1,067,290
  - **ובנוסף: רכיב הבודק תחיליות וסופיות**

מפיג עמימות אוטומטי

- תוכנה המפיגה באופן אוטומטי את העמימות המורפולוגית של המחרוזת בטקסט ובכלל זה חלקי דיבר
- התוכנה משתמשת ב- Hidden Markov Model (Baum et al. 1970). הרעיון המרכזי הוא למצוא את הפירוש הסביר ביותר של המשפט לאור שכיחויות נצפות בשפה.
- מקור הידע על השכיחויות הוא **קורפוס מתויג מורפולוגית**
- רמת הדיוק של מפיג העמימות האוטומטי היא 90% עבור ניתוח מורפולוגי מלא ו-95% עבור זיהוי של חלקי דיבר.

קורפוסים של NITE

קורפוס M1 (לא מתויג)

- קורפוס הכולל כמיליון מילים שהופקו מתוך 644 טקסטים שנלקחו ממקורות שונים וסוגות שונות
- הקורפוס משמש למחקר ופיתוח

קורפוס מתויג

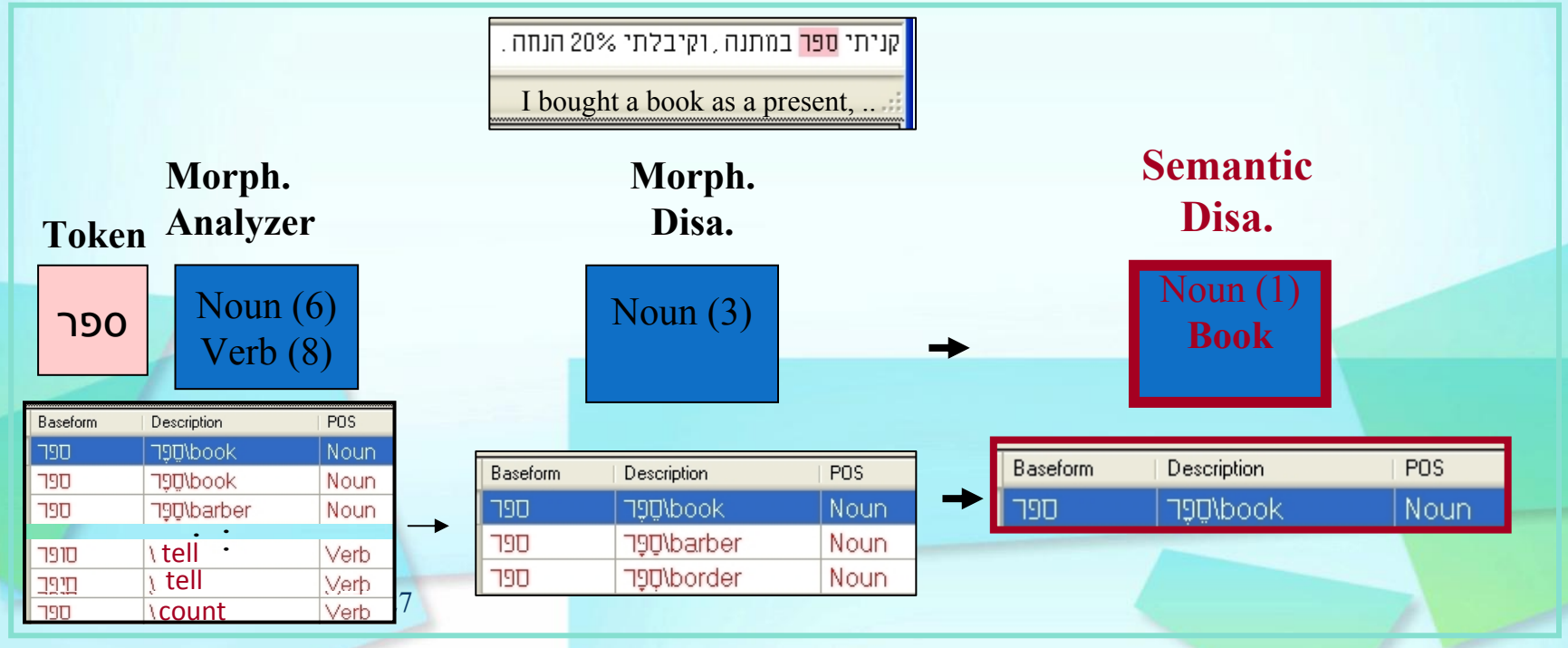
- כולל 260,000 מילה מתוך 700 יחידות טקסט בנות 500 מילה כל אחת
- כל הטקסטים תויגו על ידי לשונאי מומחה

רמת הדיוק של מפיג העמימות האוטומטי :

- 90% עבור ניתוח מורפולוגי מלא ו-95% עבור זיהוי של חלקי דיבר.

מפיג עמימות סמנטי

- רכיב תוכנה המסייע למפיג העמימות האוטומטי בהפגת עמימות סמנטית עבור מחרוזות עם ניתוח מורפולוגי זהה. בסיום שלב זה יש כבר אפשרות לזהות את הערך המילוני של המחרוזת – **הלקסמה**.



Semantic Disambiguation Algorithm

- The algorithm is based on Context-Vectors (CV). For each lexeme the lexemes which co-appear with it in the sentence.
- Training phase: CVs are built from a manually Labeled Corpus for each lexeme in the corpus.

  sentences.) 8,670 lexeme types,12,598(
- Disambiguation phase: for each ambiguous lexeme in a given sentence, the CV of the sentence is compared to the CVs of all the possible corresponding lexemes
- The accuracy rate is 95%

קניתי ספר במתנה, וקיבלתי 20% הנחה.

I bought a book as a present, ..

Sentence Context Vector

	Read	Buy	Gift	Discount	Hair	Shampoo	City	passport
	0	**1**	**1**	**1**	0	0	0	0

מתאם עם וקטור BARBER : -0.358

מתאם עם וקטור BARBER : -0.433

מתאם עם וקטור BOOK : 0.310

Corpus Context Vectors

Lexeme	Read	Buy	Gift	Discount	Hair	Shampoo	City	passport
Book-Noun	**12**	**5**	**4**	**8**	1	0	2	1
Barber-Noun	3	1	1	5	10	13	2	0
Border-Noun	2	3	1	0	0	2	10	12

שמו עשו ואוזרי כן יצא אוזיו וידו אחזת בעקב עשו ויקרא
שמו יעקב ויצחק בן ששים שנה בלדת אתם ויגדלו
הנערים ויהי עשו איש ידע ציד איש שדה ויעקב איש תם
יושב אהלים ויאהב יצחק את עשו כי ציד בפיו ורבקה
אהבת את יעקב ויזד יעקב נזיד ויבא עשו מן השדה והוא
עיף ויאמר עשו אל יעקב הלעיטני נא מן האדם האדם

כו שְׁמוֹ עֵשָׂו׃ וְאַחֲרֵי־כֵן יָצָא אָחִיו וְיָדוֹ אֹחֶזֶת בַּעֲקֵב עֵשָׂו וַיִּקְרָא
כז שְׁמוֹ יַעֲקֹב וְיִצְחָק בֶּן־שִׁשִּׁים שָׁנָה בְּלֶדֶת אֹתָם׃ וַיִּגְדְּלוּ
הַנְּעָרִים וַיְהִי עֵשָׂו אִישׁ יֹדֵעַ צַיִד אִישׁ שָׂדֶה וְיַעֲקֹב אִישׁ תָּם
כח יֹשֵׁב אֹהָלִים׃ וַיֶּאֱהַב יִצְחָק אֶת־עֵשָׂו כִּי־צַיִד בְּפִיו וְרִבְקָה
כט אֹהֶבֶת אֶת־יַעֲקֹב׃ וַיָּזֶד יַעֲקֹב נָזִיד וַיָּבֹא עֵשָׂו מִן־הַשָּׂדֶה וְהוּא
ל עָיֵף׃ וַיֹּאמֶר עֵשָׂו אֶל־יַעֲקֹב הַלְעִיטֵנִי נָא מִן־הָאָדֹם הָאָדֹם