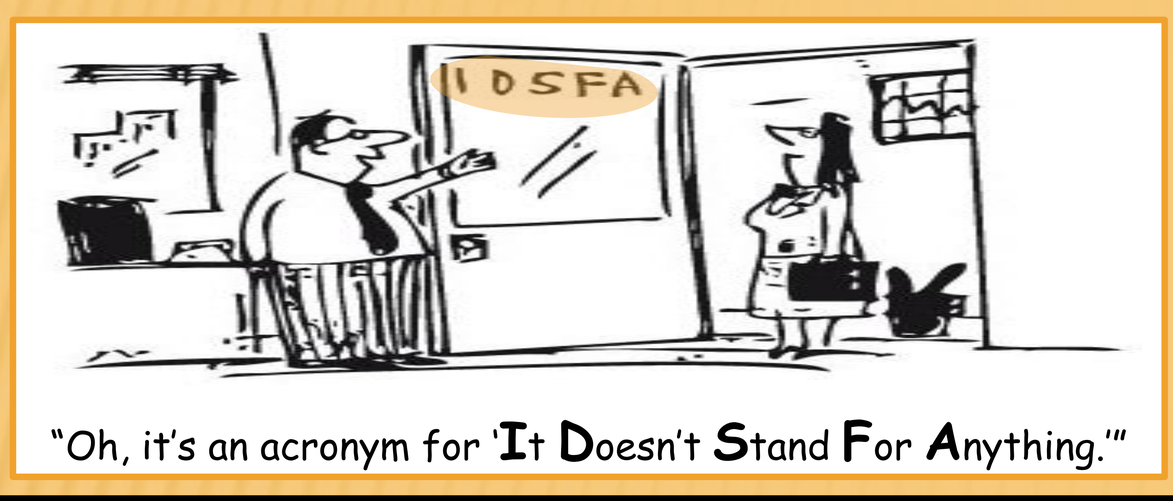
Hebrew Acronyms:
Identification, Expansion, & Disambiguation

Kayla Jacobs (Technion), **Alon Itai** (Technion), **Shuly Wintner** (University of Haifa)

Abstract

- Automatically build **acronym dictionary**
  - Rank by context match
  - Include acronyms unaccompanied by their expansions
- Improve **acronym disambiguation**
- Statistically-based **linguistic insights**



"Oh, it's an acronym for '**I**t **D**oesn't **S**tand **F**or **A**nything.'"

Why We Care

- Acronyms affect NLP applications (search, machine translation, …)
- Hand-crafted dictionaries incomplete and require constant updating.

Previous Work

- Prior acronym dictionary-building techniques rely on ***local acronyms*** (acronyms adjacent to their expansions, often in parentheses).

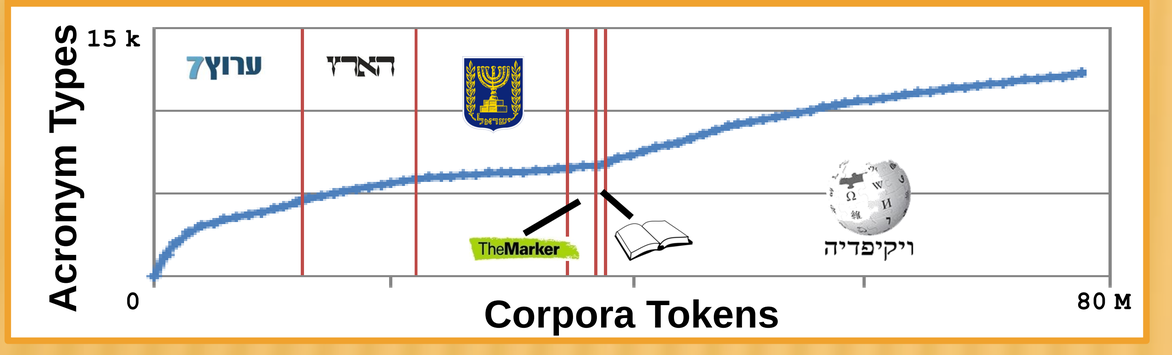
"The **CIA** (**C**entral **I**ntelligence **A**gency) released its budget."
"She works at the **C**ulinary **I**nstitute of **A**merica (**CIA**)."
"Alumni of the **C**leveland **I**nstitute of **A**rt support the **CIA**."

- Only computational work on Hebrew acronyms: HaCohen-Kerner [04,08,10,13]
  - Disambiguation of Hebrew/Aramaic acronyms in Jewish law domain.
  - Assumes a pre-existing, hand-crafted acronym dictionary.

Hebrew Acronyms

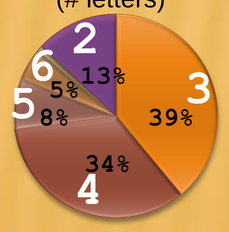
- In Hebrew corpus, acronyms are **1% of word tokens** and **3% of types**.
- More common in news and encyclopedia genres than in literature.

A never-ending story for unique acronyms:
new acronyms continue to be found as more text is read

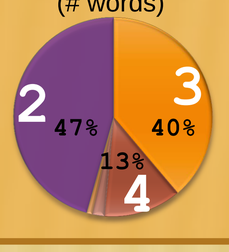


- Study **gold dictionary**: 1,000 hand-crafted acronym-expansion pairs.

Acronym Lengths
(# letters)



Expansion Lengths
(# words)



Formation Rules
(letters matched between acronym and expansion)

Letters	%	Formation Rule	Example
2	98%	□□■ □□■	ש"ח = שקל חדש
3	48%	□□■ □□■ □□■	אל"ה = איסוף למען הסביבה
	18%	□□■ □■■	ממ"ד = ממלכתי דתי
	18%	□■■ □□■	חה"א = חיל האוויר
4	21%	□□■ □□■ □□■ □□■	אעפ"כ = אף על פי כן
	18%	□■■ □■■	דוא"ל = דואר אלקטרוני
	13%	□□■ □■■■	מוצ"ש = מוצאי שבת

Building a Dictionary

1 Identify acronyms

✗ Word with internal double-quote mark ("): "הילדים וערוץ

✗ Word with double-quote mark (") before the last letter: מנכ"לית

✓ **Word of 2+ letters with double-quote mark (") before the last letter, optionally with a valid Hebrew suffix: ...ניק-, ו-, י-, ים-, ית-**

2 Identify potential expansions

- Collect corpus *n*-grams ($2 \le n \le 5$).
- Discard *n*-grams that are infrequent or end with a preposition/quantifier.

Example: מדברים על בית חולים.

n	*n*-grams	Freq.
2	✗ מדברים על	2607
	על בית	1080
	בית חולים	906
3	מדברים על בית	1 ✗
	על בית חולים	8
4	מדברים על בית חולים	1 ✗

3 Pair acronyms and *n*-grams

- For each *n*-gram, generate possible acronyms via formation rules.
- Discard infrequent acronyms.
- Tag with contextual info from LDA topic model.

Example: בית חולים

Rule	Acronym	Freq.
□□■ □□■	ב"ח	6
□□■ □■■	בי"ח	144
□■■ □■■	ביח"ו	0 ✗
□□■ □■■■	ביתח"ו	0
□■■ □■■■	ביחו"ל	0 ✗

4 Classify acronym-expansion pairs

- Train SVM to recognize matches.
- Training examples:
  - ⊕ Gold dictionary acronym paired with its gold expansion.
  - ⊖ Gold dictionary acronym paired with a non-gold *n*-gram.
- **Linguistically-motivated classification features:** corpus frequencies, formation rule, *n*-gram PMI, acronym/*n*-gram lengths, LDA topic similarity.

Example: בי"ח

Acronym	*n*-gram	
בי"ח	בית חולים	⊕
בי"ח	ביו חוות	
בי"ח	בא יחד	⊖
בי"ח	באמונתו יחיה	

Approach	Precision	Recall	F-score
Baseline Guess acronym's most-frequent *n*-gram is correct expansion	55 %	3 %	5 %
Our classifier	**82 %**	**81 %**	**82 %**

Acronym Disambiguation *(Extrinsic Eval)*

- Given 200 acronyms and their contexts, how many of the *correct* expansions are in the top *r* dictionary results for the acronyms?

Dictionary	*r* = 1	*r* = 2	*r* = 3	*r* = ∞
Baseline #1: Dictionary of local parenthetical acronyms				52 %
Baseline #2: Gold dictionary	66 %	77 %	78 %	83 %
Our dictionary	**73 %**	**79 %**	**81 %**	**85 %**

Error Rate Reduction	*r* = 1	*r* = 2	*r* = 3	*r* = ∞
Our Dictionary *vs.* Baseline #1				**69 %**
Our Dictionary *vs.* Baseline #2	**18 %**	**8 %**	**14 %**	**14 %**

Future Work

- Try **specialized Hebrew genres/domains**: military, Jewish legal texts.
- Apply to **other languages**.
  - *Hebrew advantages:* Easy acronym identification, widespread acronym use.
  - *Hebrew disadvantages:* Complex morphology/orthography, poor NLP resources.
- **Additional applications**: search, machine translation.
  - Exploit for multi-word expressions (MWEs) and named entity recognition (NER).