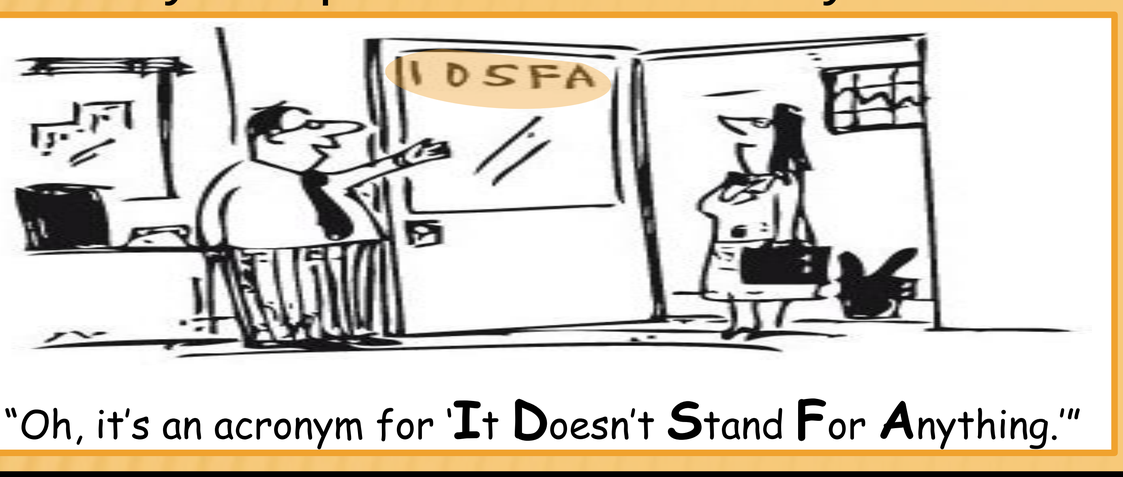
Acronyms:

Dictionary Construction & Disambiguation

Kayla Jacobs (Technion), **Alon Itai** (Technion), **Shuly Wintner** (University of Haifa). [WG1]

Abstract

- Automatically build **acronym dictionary**
  - Apply to Hebrew
  - Rank multiple expansions by context match
  - Include local acronyms (unaccompanied by expansions)
- Improve **acronym disambiguation**
- Acronym expansions are usually **MWEs**



"Oh, it's an acronym for 'It Doesn't Stand For Anything.'"

Why We Care

- Most acronym expansions are multi-word expressions (MWEs).
- Acronyms affect NLP applications like search and 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).

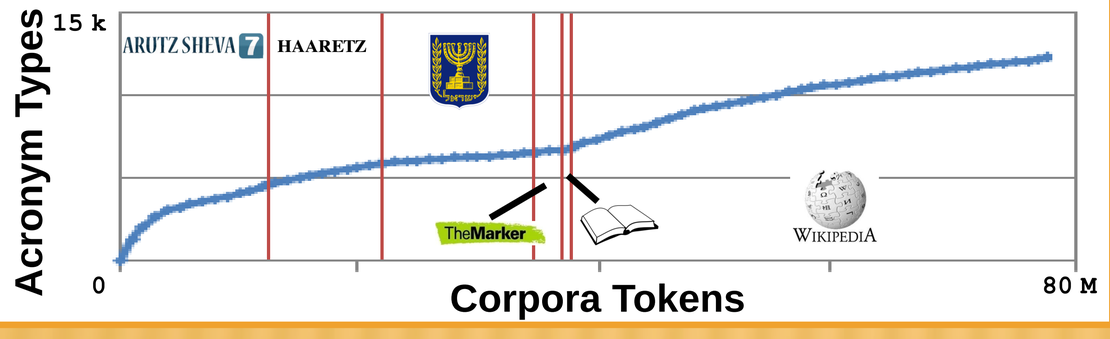
"The **Central Intelligence Agency** (**CIA**) released its budget."
"She works at the **Culinary Institute of America** (**CIA**)."
"Alumni of the **Cleveland Institute of Art** support the **CIA**."

- 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.

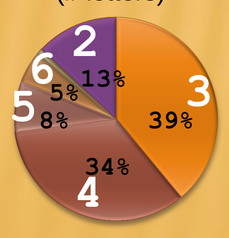
Hebrew Acronyms

- In Hebrew corpus, acronyms **1% of word tokens** and **3% of types.**
- More common in news and encyclopedia genres than in literature.
- 2000+ years of frequent usage in Hebrew; ~100 years in English.
- Challenges from Hebrew's complex morphology and orthography.

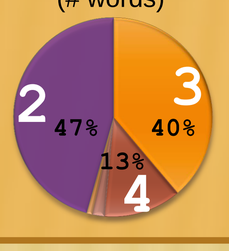
A never-ending story for unique acronyms:
new acronyms continue to be found as more text is read



Acronym Lengths
(# letters)



Expansion Lengths
(# words)



Letters	%	Formation Rule	Example
2	98%	■□□ ■□□	ש"ח = שֶׁקֶל חָדָש (*shekel new,* "New Israeli Shekel")
3	48%	■□□ ■□□ ■□□	אא"כ = אֶלָּא אִם כֵּן (*but if thus*, "unless")
	18%	■■□ ■□□	בי"ח = בֵּית חוֹלִים (*house-of sick-people*, "hospital")
	18%	■□□ ■■□	מו"מ = מַשָּׂא וּמַתָּן (*take and+give*, "negotiation")
4	21%	■□□ ■□□ ■□□ ■□□	אעפ"כ = אַף עַל פִּי כֵן (yet on as thus, "nevertheless")
	18%	■■□ ■■□	דוא"ל = דוֹאַר אֶלֶקטרוֹנִי (mail electronic, "e-mail")
	13%	■■■□ ■□□	מוצ"ש = מוֹצָאֵי שַׁבָּת (exits-of Sabbath, "Saturday night")

Building a Dictionary

1 Identify acronyms

- **Easy in Hebrew:** unambiguous orthographic marking (internal " mark).
  - יו"ר = יוֹשֵׁב רֹאש (*sitter head*, "chairperson")
- **Difficult in English:** capitalization and punctuation vary widely:
  - **M.S.** / **MS** / **M.Sc.** / **MSc** / **MSC** = **M**aster of **Sc**ience
  - **au** = **a**tomic **u**nit

2 Identify potential expansions

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

Public relations is easy.

n	*n*-grams	Freq.
2	public relations	1092
	relations is ✗	152
	is easy	5224
3	public relations is ✗	102
	relations is easy	23
4	public relations is easy	1 ✗

3 Pair acronyms and expansions

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

public relations

Rule	Acronym	Freq.
■□□ ■□□	**PR**	5293
■□□ ■■□	**PRE**	2 ✗
■■□ ■■□	**PURE**	53

4 Train classifier

- Train SVM to recognize matches:
  - ⊕ Pairs from gold dictionary
  - ⊖ Gold dictionary acronym paired with non-gold *n*-gram

Acronym	*n*-gram	
PR	public relations	⊕
PR	prince reacted	
PR	positive result	
PR	past race	⊖

- **Linguistically-motivated classification features:**
  n-gram PMI ▪ acronym and *n*-gram document frequencies ▪ formation rule acronym and *n*-gram lengths ▪ LDA topic similarity score

Match-Recognition 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

- Extrinsically evaluated dictionary on acronym disambiguation task.
- 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 %	78 %	81 %	83 %
Our dictionary	**72 %**	**79 %**	**77 %**	**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

- Exploit for identifying **multi-word expressions (MWEs)**.
- Apply to **other languages**
  - *Hebrew advantages:* Easy acronym identification, very widespread acronym use.
  - *Hebrew disadvantages:* Complex morphology/orthography, poor NLP resources.
- **Additional applications**: search and machine translation.