# Hebrew Acronyms: Identification, Expansion, and Disambiguation

**Kayla Jacobs**

Technion - Computer Science Department - M.Sc. Thesis MSC-2014-13 - 2014

# Hebrew Acronyms: Identification, Expansion, and Disambiguation

Research Thesis

Submitted in partial fulfillment of the requirements
for the degree of Master of Science in Computer Science

**Kayla Jacobs**

Submitted to the Senate of
the Technion — Israel Institute of Technology
Tishrei 5775 Haifa October 2014

The research thesis was done under the supervision of Prof. Alon Itai of the Technion Computer Science Department and Prof. Shuly Wintner of the University of Haifa Computer Science Department.

**Acknowledgments:**

My heartfelt thanks go first to my superb advisors, Alon Itai and Shuly Wintner. It was on but a whim that I first registered for Alon's "Introduction to Natural Language Processing" course, but it didn't take long before NLP (and Alon, and soon Shuly) had a devoted new enthusiast. In addition to being deeply knowledgeable and experienced, Alon and Shuly were unfailingly supportive, kind, and encouraging. Under their direction, I grew from novice grad student student to confident researcher. I already miss our vigorous research debates (including fascinating meeting detours about linguistic trivia), which were always the highlight of my week. I am so very grateful to have found such true teachers, generous guides, and motivating mentors. Thank you so very much!

I am fortunate to have learned from many others as well. My machine learning horizons were wonderfully expanded by Ran El-Yaniv (who also, along with Nachum Dershowitz, provided lively discussion at my thesis defense), Doug Freud, Rayid Ghani, Assaf Glazer (several times over), Kwang-Sung Jun, Shie Mannor, and Shaul Markovitch. Rafi Cohen kindly introduced me to LDA, and Yulia Tsvetkov to LSA. Tony Rieser indulged me mathematically, and my calculations received statistically significant improvements from Nicholas Mader, Breanna Miller, Zach Seeskin, and Brandon Willard. And in preparing me for both the pleasures and rigors of research, Suzanne Flynn, Nili Sadovnik, and Jeff Holcomb stand out as educational inspirations.

Friends made graduate school so much sweeter. My adorable officemate Limor Leibovich was always ready to greet me with a smile (and a gentle correction of my ever-improving Hebrew vocabulary). Other fellow Technion student partners-in-crime Yosi Atia, Hanna Fadida, Daniel Hurwitz, Tova Krakauer, and Arielle Sullum all helped keep me sane (and fed) during our wonderful campus lunch breaks. Aparna Rolfe, Matt Steele, and Dan Zaharopol may have been a bit too far for lunch after I moved to Israel, but they were never too far for enduring friendship. Lior Leibovich, Shachar Maidenbaum, Elisheva Rotman, and especially Beny Shlevich cheerfully offered free annotation and translation help in addition to their fond friendship.

Warm thanks go to my family. My father, Avrom Jacobs, and brother, Gilad Jacobs, were constant cheerleaders, always lovingly boasting about how little they understand my work's technical details. Phyllis and Tommy Koenigsberg have loyally served as honorary grandparents ever since I arrived in Israel, never missing a Friday afternoon call. My dear, delightful "Haifa moms" and their welcoming families, especially the Gershons and Pinskys, have been truly tremendous examples of hospitality and warmth.

Most of all, my husband, Chaim Kutnicki, quietly contributed so much patience, loyalty, perspective, more patience, support, humor, even more patience, code that worked much more efficiently than mine, and turtles. (I could not have done this without you, my love.)

Finally, I dedicate this work to my beloved mother, Dr. Dr. L. F. Jacobs, B.S., B.S., M.S., Ph.D., M.D., ז"ל, whose acronym-ful name was perhaps an inspiration for this thesis, and whose life was definitely an inspiration for so much more.

The generous financial support of the Technion is gratefully acknowledged.

# Contents

# List of Figures

# List of Tables

# Hebrew Transliteration and Translation

To facilitate the readibility of Hebrew characters, we provide a Roman character transliteration using `typewriter font`, following the schema developed by MILA: Knowledge Center for Processing Hebrew [21]:

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| א | ב | ג | ד | ה | ו | ז | ח | ט | י | כ |
| a | b | g | d | h | w | z | x | v | i | k |
| ל | מ | נ | ס | ע | פ | צ | ק | ר | ש | ת |
| l | m | n | s | y | p | c | q | r | e | t |

Hebrew does not have upper-case and lower-case letter versions, but does have a special form for five letters when they appear at the end of a word. No distinction is made in the transliteration scheme for these final form letters: כ = ך = k; מ = ם = m; נ = ן = n; פ = ף = p; and צ = ץ = c.

Though Hebrew is read right-to-left, the transliteration is read left-to-right.

Throughout this work, we follow examples of Hebrew text with a parenthetical English explanation: first a word-by-word gloss in italics, and then an overall phrase translation in quotation marks.

# Abstract

Acronyms are words formed from the initial letters of a phrase. For example, CIA is a well-known acronym for the Central Intelligence Agency, though in other contexts could mean the Culinary Institute of America or Rome's Ciampino Airport. Understanding acronyms is important for many natural language processing applications, including search and machine translation.

While hand-crafted acronym dictionaries exist, they are limited and require frequent updates. We developed a new machine learning method to automatically build a Modern Hebrew acronym dictionary from unstructured text documents. This is the first such technique, in any language, to specifically include acronyms whose expansions do not necessarily appear in the same documents. We also enhanced the dictionary with contextual information to help select the expansions most appropriate for a given acronym in context. When applied to acronym disambiguation, our dictionary achieved better results than dictionaries built using prior techniques.

Additionally, while acronyms have a long history in Hebrew, and have previously been investigated from a linguistic perspective, they have never before been studied quantitatively. We discovered new statistically-based linguistic insights about acronym usage in Modern Hebrew texts, of interest to Hebrew language aficionados and developers of Hebrew natural language processing systems.

**Keywords:**
Hebrew Acronyms, Acronym Dictionary, Acronym Disambiguation

# Chapter 1

# Introduction

## 1.1 Basics of Acronyms and Expansions

An *acronym* is a word typically formed from the initial letters of two or more other words, called its *expansion*. For example, CIA is a well-known acronym for the Central Intelligence Agency, though it has additional possible expansions including the Culinary Institute of America and Rome's Ciampino Airport.

Acronyms are a relatively recent addition to the English language, first significantly appearing in the 20th century [26], and in recent years becoming increasingly popular in internet- and phone-based communications (e.g., LOL = laugh out loud, FAQ = frequently asked questions, BCC = blind carbon copy) [7].

By contrast, Hebrew has a long history of acronyms, dating back to the Mishnaic era of the 1st–4th centuries CE [41]. Acronyms are especially frequent in the specialized genres of Jewish religious and legal texts of all historical periods [17] and in modern Israeli military writings [41] (see Figure 1.1); overall, in the secular Modern Hebrew texts we investigated, acronyms account for about 1% of word tokens and 3% of word types[1]. Hebrew acronyms have been previously studied from a linguistic perspective, but never before from a quantitative/statistical angle.

[1] Word *tokens* are individual occurrences of words, which are made up of unique word *types*. For example, the sentence "A rose is a rose is a rose." has eight word tokens of three word types ("a," "rose," and "is"). In our work, we did not consider words with non-Hebrew characters, numerals, or punctuation to be Hebrew words.


Figure 1.1: Example of frequent acronym usage in the Israeli military, in a notice posted in an armored personnel carrier (APC). Of the 17 Hebrew tokens in the sign, six (35%) are acronyms. *Credit: Chaim Kutnicki.*

Understanding the relationship between acronyms and their expansions is important for several natural language applications, including:

- **Information retrieval:** When searching for a document using a query containing an acronym, documents containing its expansion should also be returned—and vice versa.
- **Machine translation:** When automatically translating text from one language to another, acronyms often present a challenge. If the source text includes acronyms, it is rarely sufficient to simply transliterate the acronym letters; indeed, the acronym may not even exist in both languages.
- **Acronym sense understanding / disambiguation:** An acronym in text may not be familiar to the reader (whether computer or human), leaving its meaning puzzling. Alternatively, it may have additional known expansions beyond the intended one, each of which can change the interpretation of the text. Recognizing the correct meaning of an acronym, given the context, can be critical to understanding.

Currently, processing tools typically rely on "acronym dictionaries" with entries consisting of acronyms and their expansion(s). However, the collection of acronyms is an open set, with new acronyms constantly being added for company and organization names, technical terms, etc. [26]. These dictionaries are thus far from complete and require frequent updates.

To our knowledge, all existing methods to automatically build an acronym dictionary from corpora (detailed in Section 2.1) address only *local acronyms*, those whose expansions occur somewhere in the same document, typically near the first usage and often in parentheses. For example, CIA is a local acronym, with different expansions, in each of the following sentences:

- "The Central Intelligence Agency (CIA) released its budget."
- "She's applying to the CIA (Culinary Institute of America)."
- "The acronym for Rome's Ciampano Airport is CIA."
- "After graduating from the Cleveland Institute of Art, I'm a proud CIA alumnus."

In contrast, *global acronyms* are *not* accompanied by their expansions in the same document, written with the (frequently incorrect) assumption that the reader can easily understand the acronym's intended meaning. These global acronyms present a more challenging problem.

## 1.2 Research Contributions

- **Method for building an acronym-expansion dictionary with contextual information, including global acronyms:** We developed a new machine learning method to automatically extract acronyms and their expansions from unstructured corpora, to construct a context-enhanced acronym-expansion dictionary. The approach specifically includes global acronyms, making it the first work, to our knowledge, to address this important acronym class. Dictionaries built with this method are easily updatable and can be created from, and applied to, specialized domains.
- **New Hebrew language resource:** We applied our dictionary-building method to Hebrew corpora to create a new Hebrew acronym dictionary, suitable for use in natural language processing applications. While there already exist such dictionaries, ours is larger and more comprehensive, and also includes contextual information useful for disambiguating acronym meanings in texts.

- **Hebrew acronym disambiguation:** As an extrinsic evaluation of our dictionary, we applied it to the problem of acronym disambiguation in context, and achieved superior performance compared to dictionaries built with existing methods.
- **Linguistic insights about Hebrew acronyms:** We investigated the linguistic properties of Hebrew acronyms and their usage in text from a statistical angle. These insights are of interest to linguists, Hebrew language aficionados, and developers of Hebrew natural language processing systems who want their work to apply better to acronyms.

## 1.3 Resources and Tools

Our work used large unstructured text collections (corpora), as well as two additional small structured linguistic resources and four natural language processing tools.

### 1.3.1 Corpora

We combined six corpora of free Hebrew text (see Table 1.1), consisting of news articles from various Israeli news sources (Arutz 7, HaAretz, and TheMarker), records of parliamentary proceedings (Knesset), chapters of literary books (Literature), and the text content of Hebrew Wikipedia.[2] Of note, all corpora were secular publications in Modern Hebrew and not from the genre of classic Jewish texts (though a small number of documents may discuss Jewish texts or subjects).

As expected with such diverse sources, the individual documents varied significantly in average document length, vocabulary size, subject matter, and writing style. In total, the size of the combined corpora was over 77 million Hebrew word tokens (not including numbers, punctuation, or non-Hebrew tokens), slightly over half from the Wikipedia corpus.

[2]The Literature corpus was generously provided by Justin Parry of the National Middle East Language Resource Center (NMELRC). All other corpora were from MILA: Knowledge Center for Processing Hebrew [21]. The Wikipedia corpus was helpfully pre-processed by Tomer Ashur and Sela Ferdman to remove non-textual material.

| Corpus | Documents | % | Tokens | % | Types | % |
|---|---|---|---|---|---|---|
| Arutz 7 | 92,408 | 43 | 12,507,910 | 16 | 293,205 | 31 |
| HaAretz | 27,139 | 13 | 9,453,584 | 12 | 299,358 | 31 |
| Knesset | 305 | 0.1 | 12,782,676 | 16 | 189,970 | 20 |
| Literature | 714 | 0.3 | 2,402,941 | 3 | 177,162 | 19 |
| TheMarker | 837 | 0.4 | 561,524 | 1 | 58,427 | 6 |
| Wikipedia | 94,015 | 44 | 40,069,247 | 52 | 763,444 | 80 |
| **TOTAL** | **215,418** | | **77,777,882** | | **953,594** | |

Table 1.1: Corpora documents, word tokens and word types (not including numerals, punctuation, or non-Hebrew words).

### 1.3.2 Annotated Acronym-Expansion Pairs

We randomly selected 202 of all acronym types which appeared at least five times in the corpora. For each, we selected an instance of that acronym in the corpora, along with its context (the sentence and document it appeared in). If the acronym type appeared more than once in a document, we chose the first appearance. To ensure the contexts were representative, the documents were selected from the different sub-corpus collections proportionally by length (in terms of number of word tokens) of the sub-corpus. These documents were then held out of all subsequent analysis (they constituted a negligible 193, or 0.09% of the total number of corpora documents).

Native Hebrew-speakers analyzed these acronyms by hand within their document contexts and provided the expansion as well as any prefixes or suffixes (discussed in Sections 3.5 and 3.6) to identify the "base" acronyms. At least two annotators reviewed every instance to ensure high-quality annotation; disagreements were resolved by an additional reviewer.

These pairs served as an extrinsic evaluation set (see Section 4.6.1) for analyzing the quality of the acronym-expansion dictionary. In addition, they provided a detailed sample of acronyms in text for our linguistic investigations in Chapter 3, though the sample is small enough that statistical conclusions may not comprehensively reflect general acronym behavior.

### 1.3.3 Gold-Standard Acronym-Expansion Pairs

We curated a gold-standard collection of known acronym-expansion pairs collected from three online, human-edited dictionaries.[3] We discarded acronyms and expansions which appeared fewer than five times in the corpora, to ensure that the set was representative of the acronyms and expansions present in the corpora documents.

Two dictionaries included category tags like "Economics," "People," "Law," etc. We removed entries in the "Judaism" category as they belong to a different genre of text (mostly ancient and medieval Jewish law documents, which have language usage that differs significantly from the mostly secular Modern Hebrew texts of the corpora we studied).

Lastly, we manually reviewed each of the remaining pairs to discard entries that were obviously typos or mistakes. The final high-quality set consisted of 885 acronym-expansion pairs. We used this set to train and intrinsically evaluate the dictionary-building classifier in Section 4.3.2, as well as for our linguistic investigations in Chapter 3.

### 1.3.4 Tools

We used several freely-available software tools:

- **Tokenizer:** Corpora were pre-processed from their original plain text format into a tokenized XML format, using the MILA Hebrew Tokenization Tool [21]. This format includes tagged structures denoting paragraph, sentence, and single-word token structures.
- **Morphological Analyzer:** Individual tokens were morphologically analyzed using the MILA Hebrew Morphological Analysis Tool [21]. All possible morphological analyses for each token were generated, reflecting prefixes, part of speech, transliteration, gender, number, definiteness, and possessive suffix.
- **Classifier:** We trained a dictionary-building classifier using Weka [20], a suite of open-source machine learning algorithms (see Section 4.3.2).
- **Topic Modeler:** We used the machine learning toolkit MALLET [28] for its implementation of the topic modeling algorithm of Latent Dirich-

[3] We are grateful to Josh Wortman for making one of these sets available.

let Allocation (LDA). For an introduction to topic modeling and LDA, see Appendix A.

# Chapter 2

# Related Work

## 2.1 Building an Acronym Dictionary

Almost all prior work on acronym dictionary building is for English. Some of the results are language-independent, but much is based on the particular acronym formation rules in English, which (as will be described in Section 3.4.1) differ significantly from—and are usually more complicated than—Hebrew. While a few works have looked at acronym dictionaries in other languages, such as Chinese (Fu et al. [12]), no relevant research was found for Hebrew, nor in other morphologically-rich languages which may have a more difficult multilingual combination of acronym-expansion pairs, as will be discussed in Section 3.7.1.

Schwartz and Hearst [43] created a simple approach to acronym dictionary construction, using a rule-based method for acronym recognition in which they assumed that either the acronym or the expansion is written within parentheses, such as "BLT (bacon lettuce tomato)" or "bacon lettuce tomato (BLT)." Dannélls [8] [9] expanded this algorithm and applied it to Swedish biomedical texts (one of the very few non-English examples). Park [37] also described pattern-based rules for English and identified expansions using text markers, such as parentheses and cue words (e.g., "for short"). Ji et al. [23] developed a more sophisticated English acronym-recognition regular expression and an acronym-expansion letter-matching algorithm.

A few works focused on extracting acronyms and their expansions from sources other than plain-text documents. Yi and Sundaresan [53] analyzed

web page source code, looking for HTML tags that included both an acronym and its possible expansion, such as

`<a name="CSS" href="...">Cascading Style Sheet</a>`.

Jain et al. [22] used web search query logs. They looked for consecutive queries by the same user in which first an acronym was searched for, then its (possible) expansion, following a failure of the first search query to return the desired results. For example, the first search might be for "cool," and the next for "cooperation in ontology and linguistics," providing a possible acronym-expansion pair.

Several studies (such as Zahariev [55], Dannélls [10], Xu and Huang [52], and Nadeau and Turney [33]) addressed the issue of matching and ranking potential acronyms-expansion pairs once they are identified, using machine learning and linguistically-informed features to classify pairs as related or not. We employed a similar approach in Section 4.3, albeit with some new and powerful features.

A particular specialized English domain that has received extensive acronym attention is MEDLINE, the U.S. National Institute of Health's library of biomedical research articles, which is especially rife with biomedical acronyms. Acronyms in this domain also tend to be more complicated than in non-technical English, sometimes including numerals and/or following more non-standard acronym formation rules (for example, the intimidating DNMT3B = DNA-methyltransferase 3 beta). See Schwartz and Hearst [43], Pustejovsky et al. [39], Gaudan et al. [13], and Dannélls [9].

## 2.2 Computational Approaches for Hebrew Acronyms

HaCohen-Kerner et al. [15] [16] [17] [18] developed a Hebrew and Aramaic acronym disambiguation system for classical Jewish texts, primarily in pre-Modern Hebrew. They used a pre-existing manually-crafted acronym-expansion dictionary, achieving high accuracy with machine learning techniques.

They also showed that manual acronym disambiguation in this genre was a time-consuming and difficult task for human annotators, even highly-trained domain experts given multiple-choice options which always included the correct answer [19].

To our knowledge, no other research addresses any computational or

statistical aspects of Hebrew acronyms.

## 2.3 Linguistic Properties of Hebrew Acronyms

Several studies have explored Hebrew acronyms through a linguistic lens.

Ravid [41] classified acronyms into several categories (orthographic, letter, root, stem, and contrived[1] acronyms) and demonstrated that their formation is a type of nonlinear affixation, which fits well with Hebrew's generally nonlinear structure. She noted that acronyms are typically nouns because of verb vocalization requirements, but that verbs can be derived from them by regular Hebrew rules. (Additionally, as we will discuss in Section 3.6, adjectives are derivable too.)

Tadmor [50] and Muchnik [32] discussed qualitative aspects of acronyms' formation, derivational rules, historical development, and comparisons with other languages' acronyms.

While not directly relevant to our study of written Hebrew, there is a great deal of research on phonological aspects of Hebrew acronyms (e.g., Bat-El [2], Bolozky [5], Glinert [14], Ravid [41], Tadmor [50], and Zadok [54]). A particular focus is the assumed unmarked "a" vowel sound in acronym pronunciation, which explains the much larger productivity of pronounceable acronym words in Hebrew compared to other languages, like English, that require marked vowels. Bat-El [2] investigated the grammar of Hebrew acronyms that are pronounced as words, concluding that it is the grammar of a natural language, and compared the phonological and morphological properties of acronyms to other words.

Because of the Hebrew language's long history of acronym use, there is also scholarship in Jewish studies on the role of acronyms in pre-Modern Hebrew. Spiegel [46] [48] provided a good overview, including examples of medieval rabbinic texts with acronym misunderstandings due to stylistic differences among pre-printing human copyists.

Lastly, there are several manually compiled Hebrew acronym dictionaries (e.g., Kizur [24] and Ashkenazi et al. [1]), including some for specialized genres like Hassidic and Kabbalistic texts (Stiensaltz [49]) and Biblical texts

[1] We discuss contrived acronyms, which Ravid termed "existent word acronyms," in Section 3.4.2.

(Marwick [27]). Additionally, there are general Hebrew dictionaries that include entries for acronyms (e.g., Melingo [31] and Wikimilon [51]).

# Chapter 3

# Linguistic Properties of Hebrew Acronyms

Hebrew acronyms have many interesting linguistic features, some of which we exploit for our dictionary-building research goals and some of which present especial challenges. We describe these properties and also present the results of our statistical investigations of Hebrew acronyms' linguistic phenomena. When relevant, we provide comparisons to English, the language of most prior research on acronyms.

## 3.1 Orthographic Styling

English acronyms are written in a wide variety of capitalization and punctuation styles, such as M.S. / MS / M.Sc. / MSc / MSC = Master of Science, au = atomic unit, and 3-D / 3D = 3-dimensional. This diversity of representations makes identifying English acronyms a non-trivial problem, especially because an acronym may appear in the same style as an ordinary word.

In contrast, Hebrew acronyms are easy to identify (as will be described in Section 4.1). They are almost always written as strings of two or more Hebrew letters, with an internal double-quote mark ("), called a *gershayim*, typically located before the last letter [34]. For example, mnk"l / מנכ"ל is a Hebrew acronym with the expansion mnhl klli / מנהל כללי (*manager general*, "chief executive officer (CEO)"). This makes accurate acronym identification much simpler in Hebrew, even though there are a small number

of false positives (non-acronym Hebrew words written with acronym-like orthographic styling), as will be detailed in Section 3.7.

Historically, in the pre-printing era, Hebrew acronyms were indicated through dots printed on top of each letter [47]. Even today, Hebrew acronyms are occasionally written with periods after each letter, such as `a.n.` / `א.נ.` = `adwn nkbd` / `אדון נכבד` (*sir honored*, "dear sir"). This format is generally used for historical reasons [32] or when transliterating foreign acronyms that use this style in the original language (such as English). In the corpora we studied, the number of acronyms written in this format was negligible.

Lastly, in general the five Hebrew letters `k` / `כ`, `m` / `מ`, `n` / `נ`, `p` / `פ`, and `c` / `צ` change script when positioned at the end of a word, becoming, respectively, `ך`, `ם`, `ן`, `ף` and `ץ`. This rule often applies to acronyms, such as `twrh nbiaim ktwbim` / `תורה נביאים כתובים` (*torah prophets writings*, "Bible"), which is generally written as `tn"k` / `תנ"ך` instead of `tn"k` / `תנ"כ`. However, more often, the final letter of an acronym is actually *not* written in final-form script, such as `xbr knst` / `חבר כנסת` (*member-of parliament*, "parliament member"), which is usually written as `x"k` / `ח"כ` instead of `x"k` / `ח"ך`. (This common exception is likely due to the implicit understanding that the acronym's last letter actually represents one of the *first* letters of an expansion word, where the letter is not written in final form.) Of corpora acronyms ending with one of the five relevant letters, 43% of tokens and 36% of types used final-form script versions. The 15% of acronym types which appear in the corpora in *both* versions cover 75% of the relevant acronym tokens; of these, 49% of the types had more tokens with the final-form version, 38% had more tokens with the non-final-form version, and 13% had equal numbers of tokens with each.

## 3.2 Prevalence of Acronyms in Text

Acronyms are prevalent in Hebrew texts of all kinds. In the secular Modern Hebrew corpora we studied, acronyms represent 0.98% of all word tokens (not including punctuation, numerals, or non-Hebrew words) and 2.70% of all word types. Table 3.1 shows the break-down by corpus. Note significantly lower acronym prevalence in the Literature and Wikipedia corpora compared to the parliamentary Knesset corpus and the three news corpora of Arutz 7, HaAretz, and TheMarker. These differences reflect the diverse writing styles

of the genres.

| Corpus | Acro. Tokens | % of Total Acro. Tokens | % of Corpus Word Tokens | Acro. Types | % of Total Acro. Types | % of Corpus Word Types |
|---|---|---|---|---|---|---|
| Arutz 7 | 228,633 | 30 | 1.83 | 4,731 | 37 | 3.23 |
| HaAretz | 106,415 | 14 | 1.13 | 2,806 | 22 | 1.87 |
| Knesset | 189,900 | 25 | 1.49 | 2,227 | 17 | 2.34 |
| Literature | 2,039 | 0.3 | 0.08 | 662 | 5 | 0.75 |
| TheMarker | 5,757 | 0.8 | 1.03 | 474 | 4 | 1.62 |
| Wikipedia | 233,330 | 31 | 0.58 | 9,355 | 73 | 2.45 |
| **TOTAL** | **766,074** | | **0.98** | **12,895** | | **2.70** |

Table 3.1: Acronym tokens and types in the different corpora: number of acronyms in the corpus, percentage of the entire set of acronyms included in the corpus, and the percentage of the corpus's words which are acronyms (see Table 1.1 for the corpus word tokens and types used to compute these percentages).

Table 3.2 shows the 10 most frequent acronym types, which together account for a third of all acronym tokens in the corpora. While it is not possible at this level to know with certainty the correct expansions of the acronyms for every instance they appear in the documents, we list the expansions most commonly known. The large presence of politically-oriented acronyms is due primarily to the parliamentary Knesset corpus, which uses them very frequently.

We also investigated the degree of openness of the set of acronym types. Figure 3.1 shows the continued growth in the number of acronym types as the corpora are read token-by-token, which is very similar to the growth of general (not necessarily acronym) word types in Figure 3.2.[1] Just as there are always novel words to encounter in a large corpus, so too there are always novel acronyms to encounter.

The conclusion of these figures is that despite the large size of the corpora (totaling over 77 million tokens), new acronym types continue to appear at a

[1]We also studied these growth curves with the corpus order reversed and found similar trends, ruling out corpus-specific anomalies.

| Type | % | Likely Expansion |
|---|---|---|
| hiw"r / היו"ר | 9.4 | h+iweb rae / ה+יושב ראש<br>(*the+sitter head*, "the chairperson") |
| ch"l / צה"ל | 6.6 | cba hgnh lieral / צבא הגנה לישראל<br>(*army defense for+Israel*,<br>"Israeli Defense Forces (IDF)") |
| iw"r / יו"ר | 2.8 | iweb rae / יושב ראש<br>(*sitter head*, "chairperson") |
| x"k / ח"כ | 2.5 | xbr knst / חבר כנסת<br>(*member-of parliament*, "parliament member") |
| arh"b / ארה"ב | 2.0 | arcwt hbrit / ארצות הברית<br>(*lands-of the+covenant*,<br>"United States of America (USA)") |
| d"r / ד"ר | 1.8 | dwqvwr / דוקטור<br>(*doctor*, "doctor") |
| lpnh"s / לפנה"ס | 1.7 | lpni hspirh / לפני הספירה<br>(*before the+counting*,<br>"before the common era (BCE)") |
| q"m / ק"מ | 1.5 | qilwmvr / קילומטר<br>(*kilometer*, "kilometer") |
| xd"e / חד"ש | 1.4 | hxzit hdmwqrvit lelwm wlewwiwn<br>/ החזית הדמוקרטית לשלום ולשוויון<br>(*the+front the+democratic for+peace and+for+equality*,<br>"The Democratic Front for Peace and Equality (political party)") |
| ty"l / תע"ל | 1.3 | tnwyh yrbit lhtxdewt / תנועה ערבית להתחדשות<br>(*movement Arab for+renewal*,<br>"Arab Movement for Renewal (political party)") |

Table 3.2: The 10 most frequent acronym types, with their percentage of acronym tokens. Together, they account for a third of acronym tokens.

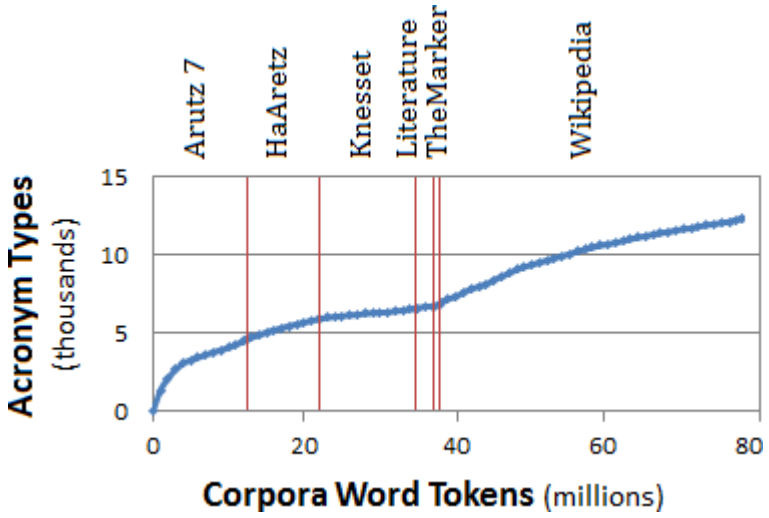



Figure 3.1: Acronym types as a function of word tokens in the corpora.

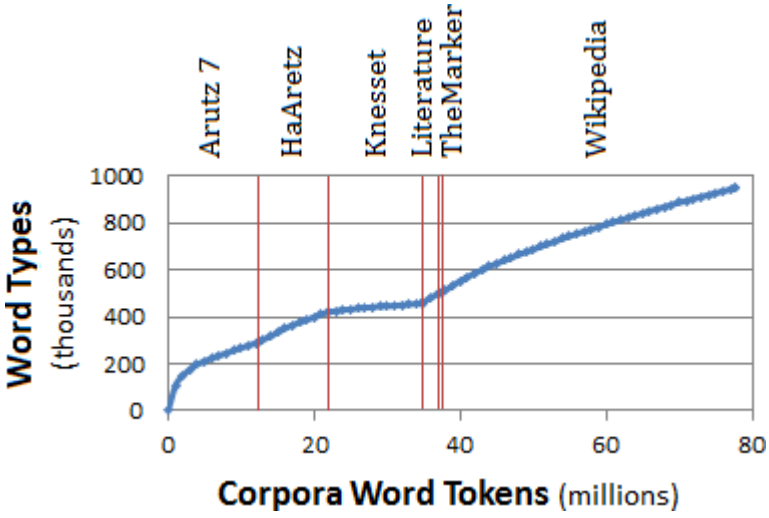



Figure 3.2: Word types as a function of word tokens in the corpora.

significant rate. This underscores the need for automated, easily-updatable methods to continue to cover the growing open set of acronyms in the language. Existing dictionaries simply can't suffice.

## 3.3 Acronym and Expansion Lengths

The lengths of acronyms and expansions in the gold-standard set are detailed in Table 3.3. Based on these statistics, we focused on acronyms of 2–6 letters and expansions of 2–5 words, which cover 99% of acronyms and expansions.

| **Acronym Length** | **%** | **Expansion Length** | **%** |
|---|---|---|---|
| 2 letters | 13 | 2 words | 46 |
| 3 letters | 39 | 3 words | 39 |
| 4 letters | 34 | 4 words | 12 |
| 5 letters | 8 | 5 words | 2 |
| 6 letters | 5 | 6+ words | 1 |
| 7+ letters | 1 | | |

Table 3.3: Lengths of (unprefixed and uninflected) acronyms and expansions.

## 3.4 Relationship Between Acronyms and Expansions

We explored several aspects of the relationship between acronyms and their expansions.

### 3.4.1 Formation Rules

An acronym is formed from its expansion by concatenating certain letters from the expansion words. The pattern of which letters are chosen is the *formation rule* for the acronym-expansion pair, and there is a strong preference for initial letters of expansion words. The most popular formation rule in both English and Hebrew takes the very first letter of each word in the expansion, such as x"k / ח"כ = x̲br k̲nst / חֲבר כְּנסת (*member-of parliament*, "parliament member"). However, we found that in over half of all Hebrew acronym types, at least one of the expansion's words contributes more than

a single letter. For example, in kdwh"a / כדוה"א = kdwr harc / כדור הארץ (*ball-of the+land*, "Earth / globe"), the first expansion word contributes the first three letters of the acronym, and the second word contributes the last two.

We introduce a notation for representing the formation rules in square brackets, with numbers denoting the position of the letter(s) of the word that appear in the acronym, and with words separated by commas. For example, the formation rule [1,1] means "concatenate the first letter of the first word, and the first letter of the second word," while [12,1,123] means "concatenate the first and second letters of the first word; the first letter of the second word; and the first, second, and third letters of the third word."

To identify the formation rules that relate acronyms and their expansions, we developed a letter-matching algorithm that, given an acronym and its expansion, outputs the formation rule(s) that relate the two. Table 3.4 shows examples of the most popular formation rules—those which account for at least 10% of the gold-standard acronym-expansion pairs of a given acronym length.

The algorithm works by matching acronym letters to the initial letters of the expansion words. It allows initial letters that could be prefixes (as will be explained in Section 3.5) to be skipped. For example, eb"s / שב"ס = eirwt bti hswhr / שירות בתי הסוהר (*service houses-of the+jailor*, "Prison Service") outputs the formation rule [1,1,h2], which skips the last word's prefix h+ / ה+ ("the"). Each rule involving skipped letters accounted for at most 1% of pairs in the gold-standard set.

Entire words can also be skipped, such as the second word in bg"c / בג"צ = bit hmepv hgbwh lcdq / בית המשפט הגבוה לצדק (*house the+law the+high for+justice*, "High Court of Justice"), which results in the formation rule [1, ,h2,l2]. However, we found the incidence of word-skipping formation rules to be negligible.

When an acronym-expansion pair was related by more than one possible formation rule, we resolved the ambiguity by choosing the rule that minimized the number of skipped words and/or letters. For example, consider the author ihwdh hlwi lwin / יהודה הלוי לוין ("Yehuda Halevi Levine"), who is often referred to by his name's acronym ihl"l / יהל"ל. The formation rule could be [1,12,1] (ihwdh hlwi lwin / יהודה הלוי לוין), or alterna-

| $m$ | Rule | Example | % |
|---|---|---|---|
| 2 | [1,1] | x"k / ח"כ = xbr knst / חבר כנסת<br>(*member-of parliament*, "parliament member") | 98 |
| 3 | [1,1,1] | aa"k / אא"כ = ala am kn / אלא אם כן<br>(*but if thus*, "unless") | 48 |
| | [12,1] | mm"d / ממ"ד = mmlkti dti / ממלכתי דתי<br>(*governmental religious*, "national religious") | 18 |
| | [1,12] | mw"m / מו"מ = mea wmtn / משא ומתן<br>(*give and+take*, "negotiation") | 18 |
| 4 | [1,1,1,1] | ayp"k / אעפ"כ = ap yl pi kn / אף על פי כן<br>(*yet on as thus*, "nevertheless") | 21 |
| | [12,12] | bim"e / בימ"ש = bit mepv / בית משפט<br>(*head-of the+government*, "prime minister") | 18 |
| | [123,1] | mwc"e / מוצ"ש = mwcai ebt / מוצאי שבת<br>(*exits-of Sabbath*, "post-Sabbath") | 13 |
| 5 | [123,12] | kdwh"a / כדוה"א = kdwr harc / כדור הארץ<br>(*ball-of the+earth*, "Earth / globe") | 20 |
| | [1,1,1,1,1] | eliv"a / שליט"א =<br>eixih lawrk imim vwbim amn /<br>שיחיה לאורך ימים טובים אמן<br>(*that+he-will-live to+length days good amen*,<br>"may he live a long good time, amen") | 13 |

Table 3.4: Examples for common formation rules in the gold-standard set. The percentages show the proportion of $m$-letter acronyms following the rule; we list all rules that are at least 10%.

tively [12,h2,1] (ihwdh hlwi lwin / יהודה הלוי לוין). However, the latter formation rule involves a skipped letter h / ה in the beginning of the second word, so it was rejected in favor of the first rule, which has no skipped letters.

Table 3.5 lists all rules that occurred in at least five pairs of the gold-standard set. Clearly, the majority of acronyms are formed by very few rules; the top seven rules cover 92% of all pairs. Note that no 5-gram formation rules appeared frequently enough to be included.

| 2-gram Rules | % | 3-gram Rules | % | 4-gram Rules | % |
|---|---|---|---|---|---|
| [1,1] | 43 | [1,1,1] | 15 | [1,1,1,1] | 2 |
| [12,1] | 13 | [1,1,12] | 1 | | |
| [1,12] | 10 | [1,1,h2] | 1 | | |
| [12,12] | 6 | [1,1,w2] | 1 | | |
| [123,1] | 3 | [1,12,1] | 1 | | |
| [123,12] | 1 | [1,h2,h2] | 1 | | |
| [1,h2] | 1 | [12,1,1] | 1 | | |
| | | [h2,12,1] | 1 | | |

Table 3.5: Formation rules which appeared with non-negligible frequency (in at least five types) in the gold-standard set, along with their proportion of the set overall.

### 3.4.2 Contrived Acronyms and Expansions

Some acronym-expansion pairs are *contrived*: the expansion (or acronym) may be deliberately designed to create an acronym (or expansion) that has an intended meaning as a word, even if this sometimes results in an awkward phrase or unusual formation rule. For example, the 2001 American law passed in response to the September 11 terrorist attacks was named the USA PATRIOT = Uniting and Strengthening America by Providing Appropriate Tools Required to Intercept and Obstruct Terrorism Act. (Unsurprisingly, the law is rarely referred to by its mouthful of an expansion.)

Similarly, for an existing expansion, the formation rule used may be deliberately chosen for the positive meaning of the resulting acronym, or to avoid an undesirable result. Consider the Israeli tutoring program for disadvantaged children, pr"x / פר"ח = prwiqv xwnkwt / פרויקט חונכות (*project-of tutoring*, "Tutoring Project"). Via the [12,1] formation rule, the acronym spells out the Hebrew word for "flower," emphasizing the program's contribution to the "blossoming" of its tutees. An alternative acronym, using the more common formation rule of [1,1], would have been p"x / פ"ח, which spells the Hebrew word for "garbage can"—not a desired association!

### 3.4.3 Orphaned Acronyms and Evolving Expansions

Some acronyms' expansions may change over time to reflect a branding shift or other evolution in meaning. For example, the American standardized test SAT originally stood for Scholastic Aptitude Test but was later officially changed to Scholastic Assessment Test. Ultimately, the SAT acronym was declared to not stand for anything at all, making it an *orphaned acronym*—still clearly identifiable as an acronym because of its all-capital letters, but no longer (officially) having an associated expansion.

Occasionally a commonly-used pronounceable acronym even gains the status of a regular, non-acronym word in its own right. For example, few English-speakers regard "laser" as anything other than a regular word, even though it originated as an acronym for Light Amplification by Stimulated Emission of Radiation. Similarly, in Hebrew, the acronym `dw"x` / **דו"ח** = `din wxebwn` / **דין וחשבון** (*judgement and+accounting*, "report") has in recent years dropped the double-quote mark to enter speakers' lexicons as the regular, pronounceable word `dwx` / **דוח** ("report"), and has even lent itself to the new verb `ldwwx` / **לדווח** (*to-report*, "to report").

### 3.4.4 Acronym and Expansion Ambiguity

A given acronym may have several possible expansions, depending on context—a phenomenon called *acronym ambiguity*. For example, the acronym `a"a` / **א"א** may mean, among other things, `alkwhwlisvim anwnimiim` / **אלכוהוליסטים אנונימיים** (*alcoholics anonymous*, "Alcoholics Anonymous (AA)"), `ai aper` / **אי אפשר** (*not possible*, "impossible"), politician `aba abn` / **אבא אבן** (*Abba Eben*, "Abba Eben"), or `anrgih avwmit` / **אנרגיה אטומית** (*energy atomic*, "atomic energy").

More rarely, several different acronyms may be formed from the same expansion. This *expansion ambiguity* tends to occur in less common expansions, whose acronyms have not yet been standardized, though it can affect well-known phrases as well; for example, `e"b` / **ש"ב** and `eb"k` / **שב"כ** are both widely-used acronyms for Israel's `eirwt hbivxwn hklli` / **שירות הביטחון הכללי** (*service-of the+security the+general*, "General Security Service").

An important implication for dictionary-building is that there is not necessarily one correct expansion for an acronym (or vice versa), but instead there can be a set of expansions with varying degrees of appropriateness

depending on the context.

In the gold-standard Hebrew acronym-expansion pairs, a large majority (79%) of the acronym types had only one expansion listed, while 14% had two expansions, and all but one of the remaining entries had between three and six expansions (the outlier was `m"a` / **מ"א**, with an extreme 10 different expansions listed). Expansion ambiguity was less common, with over 91% of expansion types corresponding to a unique acronym. Theoretically, there is no upper bound on the degree of ambiguity possible, as writers can always choose to create yet another expansion or acronym for existing acronyms or expansions, but these figures give a sense for the typical degree of ambiguity in existing dictionaries.

### 3.4.5 Relative Acronym and Expansion Frequencies

We investigated the relative frequencies in the corpora of the gold-standard acronyms and their expansions to see if a pattern emerged. For example, do frequent acronyms have frequent expansions?

We found many examples of frequent acronyms paired with infrequent expansions (e.g., `bg"c` / **בג"צ** = `bit hmepv hgbwh lcdq` / **בית המשפט הגבוה לצדק** (*house the+law the+high for+justice*, "High Court of Justice"), with an acronym collection frequency of 4,538 vs. an expansion collection frequency of 97), infrequent acronyms paired with frequent expansions (e.g., `at"a` / **את"א** = `awnibrsivt tl abib` / **אוניברסיטת תל אביב** (*university Tel Aviv*, "Tel Aviv University"), 5 vs. 1298), as well as acronyms and expansions that appeared with similar frequencies (e.g., `axh"c` / **אחה"צ** = `axr hchriim` / **אחר הצהריים** (*after the+noon*, "afternoon"), 894 vs. 943).

Overall, no significant relationship emerged between the frequencies of acronyms and their expansions.

## 3.5 Hebrew Prefixes and Function Words

Each word in an expansion usually contributes at least one letter to the acronym. A major exception are function words like "the", "of", and "to" in English; and `lmyn` / **למען** ("for") and `el` / **של** ("of") in Hebrew. These function words are often entirely skipped when forming acronyms; for example, The Association of Americans and Canadians in Israel is represented

as AACI, not TAOAACII.

In English, function words are always separated from other words by spaces and thus are easy to recognize. In Hebrew, however, many are orthographically represented as prefixes:

1. `b+` / `+ב` ("in / on")
2. `h+` / `+ה` ("the")
3. `w+` / `+ו` ("and")
4. `k+` / `+כ` ("as")
5. `l+` / `+ל` ("to / for")
6. `m+` / `+מ` ("from")
7. `e+` / `+ש` ("that")

Certain combinations of these prefixes are also possible, such as `mh+` / `+מה` ("from the") or `wke+` / `+וכש` ("and when").

Prefixes preceded slightly more than half (51%) of acronym tokens in the annotated set. Table 3.6 lists the most common ones, with comparisons to the prefixes of *word* (not necessarily acronym) tokens in the MILA Hebrew Tree-bank [45], a 6,500-sentence hand-analyzed subset of the HaAretz corpus. Note that the frequencies of `b+` / `+ב` ("in / on") and `h+` / `+ה` ("the") are quite different for the annotated set's acronyms and the tree-bank's word tokens.

One problem with prefix function words is the danger of misidentifying them as non-prefixed initial letters of acronyms. For example, `bi"d` / `בי"ד` could be the acronym for `bit din` / בית דין (*house-of judgement*, "court of law"), or the function word prefix `b+` / `+ב` ("in") followed by the isopsephic[2] acronym `i"d` / `י"ד` ("14"). This "prefix or not?" problem is not confined to acronyms but is a general issue with all Hebrew words. Typically, morphological analyzers use a lexicon of known word types to determine whether a given token is prefixed, or simply has a first letter (or letters) which *could* be function words but aren't.[3] However, this approach is more limited when applied to our task of acronym dictionary construction, as many acronyms are not in the lexicon.

In addition, prefixed particles *can* legitimately contribute to the acronym letters, even *instead* of the content words they precede. Consider `xw"l` /

[2]Section 3.7.2 will discuss isopsephic acronyms in detail.

[3]Of course, sometimes several morphological analyses are possible for the same token, prefixed and not, as illustrated by the `bi"d` / `בי"ד` example. Selecting the correct analysis depends on the context.

| Prefix | Meaning | % Prefixed Acronym Tokens | % Tree-bank Word Tokens |
|---|---|---|---|
| b+ / ב+ | "in / on" | 25 | 19 |
| h+ / ה+ | "the" | 19 | 45 |
| l+ / ל+ | "to / for" | 16 | 10 |
| w+ / ו+ | "and" | 12 | 11 |
| m+ / מ+ | "from" | 8 | 4 |
| wh+ / וה+ | "and the" | 6 | 0 |
| k+ / כ+ | "as" | 3 | 1 |
| mh+ / מה+ | "from the" | 3 | 0 |
| wl+ / ול+ | "and to / for" | 2 | 0 |
| e+ / ש+ | "that" | 1 | 9 |
| eb+ / שב+ | "as in / on" | 1 | 0 |
| wb+ / וב+ | "and in / on" | 1 | 0 |
| wmh+ / ומה+ | "and from the" | 1 | 0 |
| ke+ / כש+ | "when" | 0 | ∼0 |
| me+ / מש+ | "from that" | 0 | ∼0 |

Table 3.6: Common Hebrew function word prefixes, with their frequencies in the annotated set's acronym tokens and in the tree-bank word (not necessarily acronym) tokens.

חו"ל = xwc larc / חוץ לארץ (*outside to+(the+)land*, "abroad"), where the acronym's last letter comes from the function word prefix l+ / ל+ ("to / for"), while the content word arc / ארץ ("land") is not represented in the acronym at all. Complicating matters, this behavior is not ubiquitous: for example, in bg"c / בג"צ = bit hmepv hgbwh lcdq / בית המשפט הגבוה לצדק (*house the+law the+high for+justice*, "High Court of Justice"), the prefixes h+ / ה+ ("the") and l+ / ל+ ("to / for") don't appear in the acronym, and don't prevent the content words they precede from contributing their first letters g / ג and c / צ.

## 3.6 Hebrew Suffixes

Hebrew acronyms, like many words, can have a variety of suffixes attached to their ends.

The most common suffixes inflect for number and gender, with the nor-

mal patterns of suffixing +it / +יות for feminine,[4] +im / +ים for masculine (or mixed-gender) plural, and +wt / +ות for feminine plural. For instance, the acronym mnk"l / מנכ"ל = mnhl klli / מנהל כללי (*manager general*, "chief executive officer (CEO)") becomes mnk"lit / מנכ"לית when referring to a female CEO, even though the corresponding expansion is mnhlt kllit = מנהלת כללית. While English does not have grammatical gender, it does follow similar rules for pluralization, as in CD = certificate of deposit, which pluralizes to CDs = certificates of deposit.

Hebrew pronomial suffixes function as shortened forms of possessive personal pronouns. For example, the suffix +w / +ו is short for elw / שלו ("his"), as in mnk"lw / מנכ"לו = mnhl klli elw / מנהל כללי שלו (*manager general his*, "his chief executive officer (CEO)").

Hebrew acronyms can even have derivational suffixes, such as applying the +i / +י suffix to ch"l / צה"ל = cba hgnh lieral / צבא הגנה לישראל (*army defense for+Israel*, "Israeli Defense Forces (IDF)") to form ch"li / צה"לי ("pertaining to the IDF").

Independent of the type of Hebrew suffix, the double-quote mark still appears before the last letter of the acronym's non-inflected form [34] (e.g., a female CEO is mnk"lit / מנכ"לית, not mnkli"t / מנכלי"ת), making such inflections easier to identify. Table 3.7 lists all valid suffixes and their frequencies in the corpora's acronym types and tokens.

## 3.7 Special Classes of Acronyms and Acronym-Like Tokens

There are several classes of acronyms with special properties that present special challenges for dictionary-building. Additionally, there are some non-acronym word tokens written with acronym orthographic styling, which can create problematic false positives when trying to automatically identify true acronyms.

[4] There are two additional inflectional suffixes used historically for feminization, +t / +ת and +h / +ה, but +it / +ית is the form typically used for feminization of *new* words, including acronyms, in Modern Hebrew [30] [35] [50].

| Suffix | Meaning | % Acronym Tokens | % Acronym Types |
|---|---|---|---|
| +it / +ית | fem. | 0.16 | 0.91 |
| +t / +ת | fem. | ∼0 | 0.07 |
| +im / +ים | masc. pl. | 0.97 | 4.71 |
| +wt / +ות | fem. pl. | 0.27 | 0.51 |
| +i / +י | “mine” / adjective | 0.21 | 1.16 |
| +k / +ך | “yours” (sing.) | ∼0 | 0.02 |
| +w / +ו | “his” | ∼0 | ∼0 |
| +h / +ה | “hers” / fem. | 0.01 | 0.19 |
| +km / +כם | “yours” (masc. pl.) | 0 | 0 |
| +kn / +כן | “yours” (fem. pl.) | 0 | 0 |
| +m / +ם | “theirs” (masc. pl.) | 0.01 | 0.19 |
| +n / +ן | “theirs” (fem. pl.) | ∼0 | 0.06 |
| +inw / +ינו | pl., “ours” | ∼0 | ∼0 |
| +ik / +יך | pl., “yours” (sing.) | 0 | 0 |
| +iw / +יו | pl., “his” | ∼0 | 0.02 |
| +ih / +יה | pl., “hers” | ∼0 | 0.07 |
| +ikm / +יכם | pl., “yours” (masc. pl.) | 0 | 0 |
| +ikn / +יכן | pl., “yours” (fem. pl.) | 0 | 0 |
| +ihm / +יהם | pl., “yours” (masc. pl.) | ∼0 | 0.06 |
| +ihn / +יהן | pl., “theirs” (fem. pl.) | ∼0 | 0.02 |
| +iim / +יים | masc. pl., adjective | 0.04 | 0.27 |
| +iwt / +יות | fem. pl., adjective | 0.03 | 0.46 |
| +niq / +ניק | agent (masc.) | 0.02 | 0.19 |
| +niqit / +ניקית | agent (fem.) | ∼0 | 0.07 |
| +niqim / +ניקים | agent (masc. pl.) | 0.02 | 0.35 |
| +niqiwt / +ניקיות | agent (fem. pl.) | ∼0 | ∼0 |

Table 3.7: Hebrew suffixes, with their frequencies in the corpora’s acronym types and tokens.

### 3.7.1 Transliterated and Translated Acronyms

A small number of Hebrew acronyms are phonetic transliterations of acronyms from other languages, usually English. For example, the acronym `awp"a` / `אופ"א` is *not* formed from its Hebrew expansion `htaxdwt hkdwrgl hairwpait` / `התאחדות הכדורגל האירופאית` (*union-of the+soccer the+European*, "Union of European Football Associations"), but is instead a transliteration of the English acronym UEFA = Union of European Football Associations.

An acronym can even be a phonetic representation of the foreign acronym's *letters* themselves, often borrowing an English punctuation style of periods after each "letter." For example, the Hebrew acronym `ap.bi.aii` / `אף.בי.איי.`, pronounced "eff bee eye," is a transliteration of the English FBI = Federal Bureau of Investigation. We found the number of acronyms of this format to be negligible in the corpora.

Identifying the expansion for a transliterated acronym is impossible using standard letter-matching techniques. In our set of annotated acronym-expansion pairs, we found 5% of the acronym-like tokens to be of this class.

### 3.7.2 Isopsephy (Hebrew Numbers / Gematria)

*Isopsephy*, known in Hebrew as *gematria*, is the system of summing the numerical values of a word's individual letters (see Table 3.8) to represent an integer. Historically, this provided a convenient way to represent numbers before the widespread adoption of numeral scripts, and today's Modern Hebrew frequently uses this system for enumerating short lists (similar to the English practice of enumerating A, B, C, ...), and for dates of the Hebrew calendar. Additionally, in the related Jewish tradition of *gematria*, a word's isopsephic value has theological or mystical significance; a famous example is the number 18, considered a "lucky number" in Jewish culture, because it is the isopsephic value of the word `xi` / `חי` ("alive").

Numbers that can be represented by a single Hebrew letter are marked with a single-quote mark at the end; for example, `a'` / `א'` is 1, and `k'` / `כ'` is 20. All other numbers are typically written in an acronym-like orthographic style, with a double-quote mark before the last letter: for example, `k"a` / `כ"א` (20 + 1) is 21. While these "acronyms" do not have traditional expansions with matching letters, they are important to handle specially in our work: we found they comprise a non-negligible 16% of acronym types,

| א | ב | ג | ד | ה | ו | ז | ח | ט | י | כ |
|---|---|---|---|---|---|---|---|---|---|---|
| a | b | g | d | h | w | z | x | v | i | k |
| 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 20 |
| ל | מ | נ | ס | ע | פ | צ | ק | ר | ש | ת |
| l | m | n | s | y | p | c | q | r | e | t |
| 30 | 40 | 50 | 60 | 70 | 80 | 90 | 100 | 200 | 300 | 400 |

Table 3.8: Isopsephy values for Hebrew letters. The value for each letter in a word is added together to form a final sum.

as measured in the annotated set.

Isopsephic acronyms have two notable constraints which aid in distinguishing them from regular acronyms:

1. Letters must appear in descending Hebrew alphabetical (and, equivalently, numerical) order; thus 613 is always represented as `tri"g` / **תרי"ג** (400 + 200 + 10 + 3) and never in any other permutation such as `tgi"r` / **תגי"ר** (400 + 3 + 10 + 200).
2. The number must be written with as few letters as possible, favoring the largest possible letters; thus 613 is `tri"g` / **תרי"ג** and not `eei"g` / **ששי"ג** (300 + 300 + 10 + 13). For cultural reasons, there are two exceptions to this rule: 15 is written as `v"w` / **ט"ו** (9 + 6) and 16 as `v"z` / **ט"ז** (9 + 7), instead of the expected 10 + 5 and 10 + 6 respectively. This applies to larger numbers ending in 15 and 16 as well; for example, 416 is represented by `tv"z` / **תט"ז** (400 + 9 + 7) instead of `ti"h` / **תי"ה** (400 + 10 + 6).

Recent Hebrew calendar years are generally written modulo 5000 in this system. For example, the Hebrew calendar year 5774 (which corresponds to the Gregorian calendar years 2013–14) is written as `tey"d` / **תשע"ד** (400 + 300 + 70 + 4), which while technically equal to 774 is commonly understood to refer to 5774 instead of to the prehistorical year 773.[5] (This phenomenon

[5] An alternative style of representing large isopsephic numbers uses a single-quote mark (') to multiply the preceding letter's value by 1000. Thus, for example, 5774 is unambiguously represented as `h'tey"d` / **ה'תשע"ד** (5×1000 + 400 + 300 + 70 + 4). However, since these words contain a single-quote mark, they are not confused with true acronyms, and so are not relevant to our work.

is not specific to Hebrew; for example, in most modern English texts, the year '98 is generally understood to refer to 1998 instead of 98 CE.)

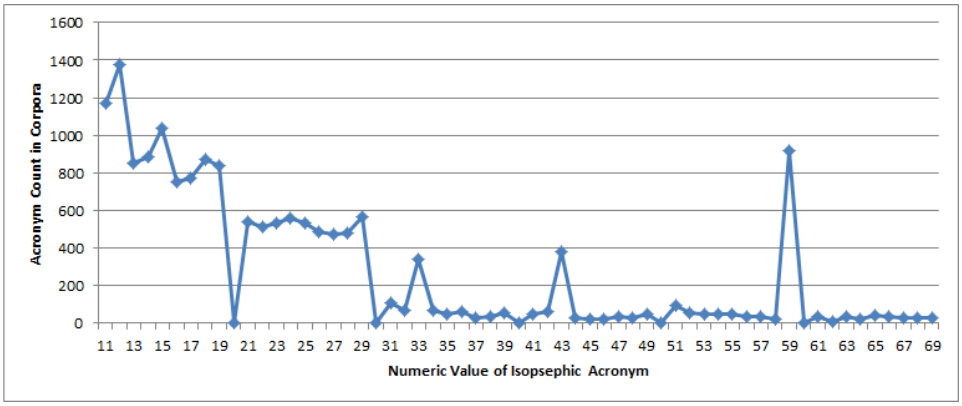



Figure 3.3: Isopsephic acronyms with numerical values 11–69.
Note much lower frequencies after 30. (The numbers 1–10, 20, 30, 40, 50, 60, are represented as single letters and not "acronyms.")

As seen in Figure 3.3, the isopsephic acronym frequency drops off sharply and suddenly, once the numerical representation exceeds 30. We hypothesize that this is due to the 30 possible days per month of the Hebrew calendar (which are usually written in isopsephic format). There is a spike shortly thereafter at 33, easily explainable by a Jewish holiday which includes this isopsephic acronym in its name. Amongst the under-30 isopsephic acronyms, the numbers 11 and 12 are by far the most popular, which is likely due to references to the 11th and 12th grades of high school, which are typically written using isopsephic forms.

Above 30, the frequency of isopsephic acronyms is usually negligible, until a steady rise from the 600s to 772 (see Figure 3.4), numbers for Hebrew calendar years which correspond to the Gregorian calendar years of about 1900–2012—i.e., the present or recent history of the corpora documents. The year 708 (`te"x` / `תש"ח`) was particularly frequently mentioned, as it is the year of the founding of the State of Israel in 1948, a major historical event for Modern Hebrew texts.

There are of course occasional frequency spikes when a non-isopsephic acronym happens to be in isopsephic-permittable form. For example, 59 is represented by `n"v` / `נ"ט`, which is also a non-isopsephic acronym with the

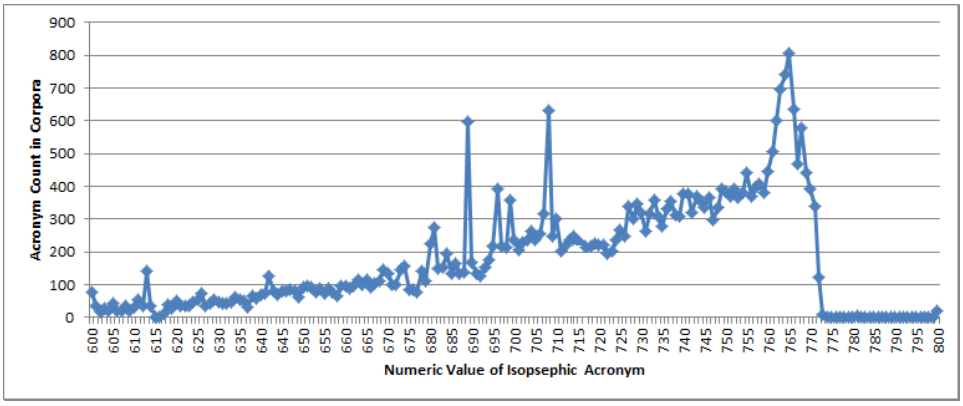



Figure 3.4: Isopsephic acronyms with numerical values 600–800, representing the Hebrew calendar years corresponding to the Gregorian calendar years 1839–2039. Note the steady rise up to 772, corresponding to 2012—the present or recent history of the corpora documents—and thereafter a sharp drop.

expansion ngd vnqim / נגד טנקים (*against tanks*, "anti-tank").

### 3.7.3 Abbreviations

*Abbreviations* are shortened words that omit certain letters or syllables, usually from the interior or end of the word. For example, in both English and Hebrew the word "professor" (prwpswr / פרופסור) is abbreviated as "prof." and prwp' / פרופ', respectively. Usually, abbreviations are marked in English with a single period at the end of the word; in Hebrew, the final punctuation mark is instead an apostrophe ('), called a *geresh*.

Because they are formed from single words, abbreviations are not true acronyms, but occasionally they are nonetheless written with an acronym-like orthographic styling, such as TV = television, d"r / ד"ר = dwqvwr / דוקטור (*doctor*, "doctor") and s"m / ס"מ = snvimvr / סנטימטר (*centimeter*, "centimeter"). Hebrew abbreviation "acronyms" are usually borrowed from other languages, such as the last two examples, which correspond to the English abbreviations "Dr." and "cm," respectively.

Thankfully, we found that in our set of annotated pairs, only 0.5% of acronym-like types were in fact abbreviations, so this phenomenon's potential for confusion is limited.

### 3.7.4 Names and Pseudonymous Initials

Some people are commonly referred to by acronyms of their names, like JFK = John Fitzgerald Kennedy.

In Hebrew, this practice is especially common when referring to rabbinical names, like the famous medieval religious commentator universally referred to as `re"i` / **רש"י** ("Rashi"), an acronym of his full name `rbi elmh icxqi` / **רבי שלמה יצחקי** ("Rabbi Solomon Yitzchaki"). In the annotated set, a full 5% of acronym types were of this class, despite the overall secular genre of the texts. Sometimes rabbinical figures are even commonly referred to by the acronym of the title of the book they are most famous for, such as the 17th century rabbi David Halevi Segel, better known as the `v"z` / **ט"ז** = `vwri zhb` / **טורי זהב** (*columns-of gold*, "Golden Columns").

Texts may also refer to anonymous or fictitious people by pseudonymous initials, which by definition have no expansion. Additionally, a few genuine names, especially last (family) names, are sometimes written in an acronym-like way. These originally may have had expansions, such as `k"c` / **כ"ץ** = `khn cdq` / **כהן צדק** (*priest holy*, "holy priest"), but they have since become orphaned acronyms with no intended expansions (as discussed in Section 3.4.3).

### 3.7.5 Spelled-Out Alphabet Letter Names

The Hebrew alphabet letters are sometimes spelled out using an acronym-like orthography: `al"p` / **אל"ף** ("aleph"), `bi"t` / **בי"ת** ("bet"), etc. In our set of annotated acronym-expansion pairs, we found only 0.5% of acronym types to belong to this class. Luckily, because this is a small and closed set of only 22 letters, they can be handled easily as special-case type additions to an acronym dictionary.

# Chapter 4

# Building an Acronym Dictionary

As described in Section 2.1, prior methods of automatic dictionary-building from unstructured texts focus on local (non-global) acronyms, whose expansions appear nearby in the same document. We developed a new method which includes global acronyms, involving three steps:

1. Identifying acronyms (Section 4.1);
2. Identifying candidate expansions (Section 4.2); and
3. Matching acronyms and expansions (Section 4.3).

The final dictionary is described in Section 4.4, with an error analysis in Section 4.5. Finally, we extrinsically evaluate the quality of the our dictionary by applying it to the problem of acronym disambiguation in Section 4.6.

## 4.1 Identifying Acronyms

To extract the set of Hebrew acronyms from the corpora, we took advantage of the unique orthography of Hebrew acronyms (described in Section 3.1). We searched for tokens that have a single internal double-quote mark, followed by either a single letter *or* a letter and then one of the valid Hebrew suffixes listed in Table 3.7.

These criteria were designed to exclude false positives of function word prefixes preceding a multiple-word quotation. For example, in `w"gbinh`

vyimh" / "וגבינה טעימה" (*and+"cheese delicious"*, "and "delicious cheese""), the word w"gbinh / וגבינה would be mistakenly identified as an acronym when simply searching for tokens with an internal double-quote mark.

The result was 12,895 acronym types, from 766,074 acronym tokens (as detailed in Table 3.1). When we removed suffixes and discarded acronym types which appeared fewer than five times in the corpora, we were left with 3,862 acronym types covering 93% of all corpora acronym tokens.

## 4.2 Identifying Candidate Expansions

To identify candidate expansions, we first extracted all $n$-grams—consecutive sequences of $n$ words—from the corpora, with $2 \leq n \leq 4$. (This range is based on the list of formation rules with non-negligible frequency, shown in Table 3.5, which covered 2-grams, 3-grams, and 4-grams, but not any $n$-grams with $n > 4$.) We discarded $n$-grams that met any of the following criteria, as they were not likely to serve as acronym expansions:

- Included any characters besides Hebrew letters (e.g., punctuation, numerals, English characters, etc.);
- Appeared fewer than five times in the corpora (as such rare $n$-grams cannot supply statistically valid information);
- Ended with a preposition or quantifier, which indicated that the $n$-gram was an incomplete phrase and thus unlikely to be a full expansion.[1]

Once we had the set of $n$-grams that might serve as expansions, we had to associate them with their possible acronyms. For every $n$-gram in the set, we generated all the acronyms that it *could* form via any of the possible formation rules in Table 3.5, and then filtered out acronyms that did not appear in the corpora at least five times.

As an illustrative example, consider the 2-gram bit xwlim / בית חולים (*house-of sick-people*, "hospital"), which appeared 906 times in the corpora—well above the threshold collection frequency of five. Table 4.1 lists the seven formation rules for 2-grams, along with the acronyms that they could generate and their collection frequencies. (Note one of the rules, [1,h2], did not

[1] Only six of the gold-standard set expansions end with a preposition or quantifier.

apply to this 2-gram because the second word, xwlim / חולים, doesn't begin with the letter h / ה as required by the rule.) Only two of these acronyms actually appeared in the corpora, however, and only one (bi"x / בי"ח) appeared above the threshold collection frequency. Thus, this $n$-gram contributed one acronym / $n$-gram pair: bi"x / בי"ח $\stackrel{?}{=}$ bit xwlim / בית חולים.

| Rule | $n$-gram | Acronym | Acronym Frequency |
|---|---|---|---|
| [1,1] | bit xwlim / בית חולים | b"x / ב"ח | 4 |
| [12,1] | bit xwlim / בית חולים | bi"x / בי"ח | 144 |
| [1,12] | bit xwlim / בית חולים | bx"w / בח"ו | 0 |
| [12,12] | bit xwlim / בית חולים | bix"w / ביח"ו | 0 |
| [123,1] | bit xwlim / בית חולים | bit"x / בית"ח | 0 |
| [123,12] | bit xwlim / בית חולים | bitx"w / ביתח"ו | 0 |
| [1,h2] | - | - | - |

Table 4.1: All acronyms formable from the 2-gram bit xwlim / בית חולים, via each of the popular formation rules for 2-grams.

After re-ordering by acronym, we had a final list of all non-rare acronyms in the corpora, each with all of its non-rare candidate expansions. Unlike the bit xwlim / בית חולים example, most $n$-grams had many possible acronym matches. Similarly, most acronyms—especially the shorter ones—had many possible $n$-gram matches, as the criteria for being considered a candidate expansion are quite inclusive. For example, the acronym bi"x / בי"ח had 640 $n$-gram pairings, a few of which are shown in Table 4.2.

## 4.3 Matching Acronyms and Candidate Expansions

Armed with the list of acronyms and their candidate expansions, the next step was to determine which ones are likely to be correct. After characterizing them by various linguistic features (Section 4.3.1), we employed standard machine learning techniques to train a classifier to distinguish between pairings of acronyms and their expansions vs. pairings of acronyms and non-expansion $n$-grams (Section 4.3.2). As will be discussed further, we used the gold-standard set for positive training examples, and also constructed a similarly-sized set of negative training examples.

| Rule | Acronym | $n$-gram |
|---|---|---|
| [1,12] | bi"x / בי"ח | ba ixd / בא יחד<br>(*come together*, "come together") |
| [1,12] | bi"x / בי"ח | bamwntw ixih / באמונתו יחיה<br>(*in+faith+his he-will-live*, "in his faith he will live") |
| [12,1] | bi"x / בי"ח | bin xwwt / בין חוות<br>(*between farms*, "between farms") |
| [12,1] | bi"x / בי"ח | bit xwlim / בית חולים<br>(*house-of sick-people*, "hospital") |
| [1,h2,h2] | bi"x / בי"ח | byiwt hihdwt hxrdit / בעיות היהדות החרדית<br>(*problems the+Judaism the+Orthodox*, "problems of Orthodox Judaism") |

Table 4.2: A few of the candidate expansions for the acronym bi"x / בי"ח.

We acknowledge a challenge to our approach: the true "correctness" of a match between an acronym and its expansion is a subjective human decision, and is a somewhat fuzzy concept to expect a machine learning classifier to tackle. Some matches are easy for humans to classify definitively (e.g., an acronym that does not match letters with the expansion is almost certainly not a match, or a commonly-used acronym and expansion that any native speaker would easily recognize as a true match), while others are more ambiguous. Additionally, the positive training examples, by dint of their source in the gold dictionaries, are generally frequent acronyms and their expansions, which could potentially behave differently than the rarer acronyms that we also included. Thus, the best a classifier can do is as well as a team of human experts—which is, however, the exact inclusion criterion for existing human-curated dictionaries.

### 4.3.1 Classification Features

For each pairing of an acronym and $n$-gram, we calculated a 44-dimensional feature vector containing measures of various linguistically-motivated properties of the acronym, $n$-gram, or relationship between them.

**Pointwise Mutual Information (PMI) of $n$-gram** Pointwise mutual in-

formation (PMI) is an association measure that quantifies the degree of collocability of words. Intuitively, a high PMI indicates that the words in the $n$-gram appear together more frequently than would be expected from their independent frequencies alone, regardless of whether the individual words appear frequently or not in text. For example, the words "new" and "race" appear relatively frequently in English texts, while the word "york" is rare; however, the bigram "new york" has a high PMI and "new race" has a low PMI.

The PMI of a bigram $AB$ is defined as:

$$\mathrm{PMI}(AB) \equiv \log \frac{p(AB)}{p(A) \cdot p(B)}. \tag{4.1}$$

The probability $p(X)$ is estimated from the corpora, and defined by:

$$p(X) \equiv \frac{\#X}{N} \tag{4.2}$$

where $\#X$ is the number of occurrences of type $X$ in the corpora, whose size (in tokens) is $N$.

Since our candidate expansions are between 2 and 5 words in length, for $n$-grams $ABC...Z$, with $n > 2$, we used an extension of the bigram PMI formula:

$$\mathrm{PMI}(ABC...Z) \equiv \log \frac{p(ABC...Z)}{p(A) \cdot p(B) \cdot ... \cdot p(Z)}. \tag{4.3}$$

We hypothesized that $n$-grams with high PMI values are more likely to be acronym expansions than $n$-grams with low PMI, as they are more likely to be phrasal units for which having a shorter acronym is useful.

**Formation Rule** We encoded the particular formation rule that described the letters of the $n$-gram that were used to match it with the acronym. We also included as features the rule's popularity (via the percentage of acronyms of the same length which are formed by this rule, as illustrated in Table 3.4), and the rule's ranking ("1" for most popular, "2" for second-most popular, etc.), reasoning that correct matches were more likely with more popular formation rules.

**Acronym and $n$-gram Lengths** We included features indicating the num-

ber of letters in an acronym and the number of words in the $n$-gram, as well as the ratio between them. As detailed in Section 3.3, certain acronym and expansion lengths are more common than others.

**Acronym and $n$-gram Collection and Document Frequencies** For each acronym and for each $n$-gram, we counted their collection frequencies (number of times they appeared in the corpora).

A related feature, the inverse document frequency (IDF), is a popular way to measure whether the acronym or $n$-gram is common or rare across all documents [36]. It is defined, for type $X$ and document set $D$, by:

$$\mathrm{IDF}(X) \equiv \log \frac{|D|}{|d \in D : X \in d|}. \tag{4.4}$$

For both the collection and document frequencies, values were calculated with respect to each of the six individual corpora, as well as for the collection as a whole, to allow for the possibility of corpus-specific behavior.

**LDA Topic Similarity of Acronym and $n$-gram** Latent Dirichlet Allocation (LDA)[2] is a topic modeling algorithm which discovers hidden (latent) themes in large textual datasets. We used LDA to model topics in the corpora, to capitalize on the intuition that acronyms and their expansions tend to appear in similarly-themed document contexts. For example, if the acronym `bi"x` / `בי"ח` appears strongly in healthcare-related documents yet weakly in art-related documents, so too ought its expansion, `bit xwlim` / `בית חולים` (*house-of sick-people*, "hospital")—but not likely its other matched $n$-grams.

To formalize this observation, we computed features representing the degree of topic similarity between the acronym and its paired $n$-gram. We first built, from the corpora, an LDA model with $T = 300$ topics. We represented the acronym as a vector $\overrightarrow{a} = (a_1, a_2, ..., a_T)$ over the topic space, where coordinate $a_i$ is the acronym's score for topic $i$ as given by the LDA model.

Similarly, the $n$-gram was represented as a vector $\overrightarrow{e} = (e_1, e_2, ..., e_T)$

[2]For more background on topic modeling and LDA, including how we determined the number of topics in the model, see Appendix A. LDA was developed in 2003 by David Blei, Andrew Ng and Michael Jordan.

where $e_i$ is the $n$-gram's score for topic $i$. Determining the coordinate values for the $e_i$'s were less obvious, as the LDA model provided topic scores for individual tokens, not multi-word $n$-grams. We therefore inferred the $e_i$'s from the topic scores for the *individual tokens* of the $n$-gram in three simple ways:

1. Pointwise multiplication of the individual tokens' scores for topic $i$;
2. Pointwise addition of the individual tokens' scores for topic $i$; and
3. Pointwise addition of the individual tokens' scores for topic $i$, but with a special case ensuring a value of 0 if any of the summands are 0.

For each method of calculating $\overrightarrow{e}$, we then computed the measure of topic similarity between the acronym and the $n$-gram by taking the cosine similarity of the two vectors $\overrightarrow{a}$ and $\overrightarrow{e}$:

$$\text{TopicSimilarity}(\overrightarrow{a}, \overrightarrow{e}) = \frac{\overrightarrow{a} \cdot \overrightarrow{e}}{|\overrightarrow{a}| \cdot |\overrightarrow{e}|}. \tag{4.5}$$

Finally, these three topic similarity measures were included as classification features, representing the degree of LDA topic overlap between the acronym and the $n$-gram.

### 4.3.2 Classifier Training and Intrinsic Evaluation

We trained a binary classifier to predict, for a pairing of an acronym and an $n$-gram, whether the $n$-gram is a true expansion for the acronym.

For *positive* training examples, we used the natural source of the 885 entries on the list of pairings which happen to be part of the gold-standard set. In other words, the positive training examples were the entries from human-edited acronym dictionaries in which the acronym and expansion each appeared at least five times in the corpora, and were related by one of the common formation rules described in Section 3.4.1 (which were learned from the very same data).

For machine learning purposes, it was important to have *negative* training examples as well, ideally near misses and the same number as positive training examples. (We considered performing 1-class classification

instead—using only positive examples—but such algorithms generally perform less well than binary classifiers.)

Since there was no obvious choice for negative examples, we constructed an artificial set. We paired acronyms in the gold-standard set to $n$-grams that were *not* listed in the gold-standard set as the "correct" expansions. For example, consider the acronym bi"x / בי"ח. As shown in Table 4.2, it was paired with 640 possible $n$-grams. Only one, bit xwlim / בית חולים (*house-of sick-people*, "hospital"), was a "correct" expansion in the gold-standard set, so we designated the pair bi"x / בי"ח $\stackrel{?}{=}$ bit xwlim / בית חולים as a positive training example. The acronym was then paired with one of its remaining $n$-grams, randomly selected from the remaining list—say, ba ixd / בא יחד (*come together*, "come together")—as a negative training example: bi"x / בי"ח $\neq$ ba ixd / בא יחד. This resulted in 883 negative training examples (slightly fewer than the number of positive training examples due to two cases of acronyms with only a possible gold $n$-gram pairing and no non-gold $n$-gram pairings).

On the 1768 total training examples, with the classification features described in Section 4.3.1, we trained a support vector machine (SVM) with a linear kernel as implemented through John Platt's sequential minimal optimization algorithm (SMO) [38].[3]

For baseline comparisons, we also built two naïve classifiers:

- *Baseline #1:* A simple classifier which selects the highest-frequency $n$-gram paired with the acronym as the expansion and rejects all other $n$-grams for the acronym; and
- *Baseline #2:* A classifier identical to ours (SMO SVM with linear kernel, trained on the same set of training examples), but with only the PMI feature.

For both SVMs, performance was evaluated using 10-fold cross-validation. As shown in Table 4.3, our classifier easily outperformed the baselines.

[3] We also tried other classifiers, including:

- SVM algorithms other than SMO (notably LibSVM, developed by Chih-Chung Chang and Chih-Jen Lin [6]);
- SVMs with nonlinear kernels; and
- decision trees (the J48 algorithm, based on Ross Quinlan's C4.5 algorithm [40]).

However, these other classifiers performed worse than the SMO SVM with linear kernel.

| Approach | Precision | Recall | F-score |
|---|---|---|---|
| *Baseline #1:* Acronym's most frequent $n$-gram | 0.55 | 0.03 | 0.05 |
| *Baseline #2:* SVM with linear kernel (SMO) trained only on PMI feature | 0.61 | 0.59 | 0.60 |
| SVM with linear kernel (SMO) trained on full feature set | 0.82 | 0.81 | 0.82 |

Table 4.3: Classifier performance, trained on positive and negative acronym-expansion pair examples, compared to the baselines of predicting the most frequent expansion candidate for an acronym and a classifier trained only on the PMI feature.

Because our method of constructing negative examples involved an element of chance, we repeated it ten separate times, training otherwise-identical SMO SVM classifiers on the different training sets (though with identical positive training examples). We found the standard deviation to be just 0.83% for precision and 0.81% for recall; such low numbers indicated high robustness for our method of constructing negative examples.

Our SVM classifier gave us some interesting insights into which features were most influential in its decision-making, corresponding to which features were most prominently weighted in its calculations. Of the 44 features, the inverse document frequency and PMI of the $n$-gram were the strongest features (which motivated our designs for the baseline classifiers, making them as strong as possible while remaining simple). The three next-strongest features were the ratios of the inverse document frequencies of the acronym and $n$-gram in the Wikipedia corpus, HaAretz corpus, and total corpora, respectively. Following closely was the LDA topic similarity measure calculated using the pointwise addition method of calculating the $n$-grams' topic scores.

We further explored the impact of the LDA topic similarity features on the classifier's performance. The other methods of calculating the $n$-gram topic scores proved less effective than pointwise addition, as indicated by their relative weights in the SVM. Additionally, Table 4.4 shows that while holding out the LDA features negatively impacted performance by only a few

percentage points, training the classifier on *only* the LDA features achieved reasonably good (if still lower) performance, indicating that a great deal of useful information is contained within the LDA topic similarity scores.

| Features | Precision | Recall | F-score |
|---|---|---|---|
| All features | 0.82 | 0.81 | 0.82 |
| All features *except* LDA similarities | 0.79 | 0.79 | 0.79 |
| *Only* LDA similarity features | 0.70 | 0.69 | 0.70 |

Table 4.4: Importance of LDA features in classifier performance.

## 4.4 The Final Dictionary

The final dictionary entries came from four sources:

- **Classifier:** All acronym / $n$-gram pairs that the classifier classified as an acronym / expansion.
- **Gold:** The contents of the three manually-compiled acronym-expansion dictionaries used to create the gold-standard set (see Section 1.3.3), including pairs that were not included in that set—pairs that did not pass the frequency test, or were not related by a common formation rule.
- **Isopsephic:** All possible isopsephic acronyms (see Section 3.7.2), along with their numerical values "expansions," in ranges corresponding to relatively low numbers which could be used for enumeration, plus Hebrew calendar years from the beginning of the Hebrew calendar until the next century. These are easily generated. No isopsephic "expansions" were candidates for the classifier, as there is no typical letter-matching between them and their acronyms.
- **Special:** A few special acronym-expansion entries such as the 22 spelled-out alphabet letter names (described in Section 3.7.5).

Some examples of entries from the final dictionary are shown in Table 4.5.

Each entry was tagged with meta-data indicating its source and, when available, statistical information with respect to the corpora. This included

| Acronym | Expansion | Source |
|---|---|---|
| i"g / י"ג | 13 | Isopsephic |
| | ie gwrsim / יֵש גּוֹרְסִים<br>(*there-exist holders-of-views*, "some say") | Gold |
| | ieral glili / יִשְׂרָאֵל גְּלִילִי<br>(*Yisrael Galili*, "Yisrael Galili (name)") | Classifier |
| | iwtr grwyh / יוֹתֵר גְּרוֹעָה<br>(*more terrible*, "worse") | Classifier |
| aa"g / אא"ג | ap awzn grwn / אַף אֹזֶן גָּרוֹן<br>(*nose ear throat*, "ear, nose and throat") | Classifier, Gold |
| | andrh agasi / אַנְדְרֶה אַגַּאסִי<br>(*Andre Agassi*, "Andre Agassi (name)") | Classifier |
| e"b / ש"ב | eirwt hbivxwn hklli / שֵׁירוּת הַבִּיטָחוֹן הַכְּלָלִי<br>(*service-of the+security the+general*, "Shin Bet (Israeli security agency)") | Gold |
| | eirwt bivxwn / שֵׁירוּת בִּיטָחוֹן<br>(*service-of security*, "security service") | Classifier |
| | elwm bit / שָׁלוֹם בַּיִת<br>(*peace house*, "domestic tranquility") | Classifier, Gold |
| | emwal b' / שְׁמוּאֵל ב'<br>(*Samuel 2*, "II Samuel (Biblical book)") | Classifier, Gold |
| | eiywri bit / שִׁיעוּרֵי בַּיִת<br>(*lessons-of home*, "homework") | Classifier, Gold |
| | ebt birwelim / שַׁבָּת בִּירוּשָׁלַיִם<br>(*Sabbath in+Jerusalem*, "Sabbath in Jerusalem") | Classifier |
| plm"x / פלמ"ח | plwgwt mxc / פְּלוּגוֹת מָחַץ<br>(*forces strike*, "Palmach (underground army)") | Classifier, Gold |
| | pnsiwni lmwrim xiilim / פֶּנְסִיוֹנִי לְמוֹרִים חַיָּלִים<br>(*pensions-related to+teachers soldiers*, "pensions for soldier-teachers") | Classifier |

Table 4.5: Example entries from the final dictionary.

LDA topic scores which (as we will see in Section 4.6) is useful for contextual disambiguation. Note that entries could have multiple sources, such as being from both the Classifier and Gold sources.

## 4.5 Error Analysis

We discuss the example entries in Table 4.5 as illustrations of common errors and successes in our dictionary.

All entries from the gold and isopsephic sources are, unsurprisingly, recognizably correct expansions for their acronyms (of course, whether or not they are *the* correct expansions depends on the context in which the acronym is used, as will be explored in Section 4.6). Many, though of course not all, of the gold-source entries were also covered by the classifier. A gold example *not* covered by the classifier is `i"g` / **י"ג** = `ie gwrsim` / **יֵשׁ גּוֹרְסִים** (*there-exist holders-of-views*, "some say"), which typically appears in classical religious Jewish texts (unlike the secular Modern Hebrew corpora studied).

The first two expansions for the `e"b` / **ש"ב** acronym are especially interesting. The classifier missed the gold expansion `eirwt hbivxwn hklli` / **שֵׁירוּת הַבִּיטָחוֹן הַכְּלָלִי** (*service-of the+security the+general*, "Shin Bet (Israeli security agency)"), because it didn't consider expansions following word-skipping formation rules like this pair's `[1,h1, ]`. However, the classifier *did* include the related partial expansion `eirwt bivxwn` / **שֵׁירוּת בִּיטָחוֹן** (*service-of security*, "security service"), which followed the more common formation rule `[1,1]`, likely because of the strong topic association between the two. Elsewhere, both the classifier and the gold dictionaries also provided the expansion's other common acronym, `eb"k` / **שב"כ** = `eirwt hbivxwn hklli` / **שֵׁירוּת הַבִּיטָחוֹן הַכְּלָלִי** (*service-of the+security the+general*, "Shin Bet (Israeli security agency)").

The first classifier-sourced expansion for `i"g` / **י"ג**, the name of politician `ieral glili` / **יִשְׂרָאֵל גָּלִילִי** (*Yisrael Galili*, "Yisrael Galili"), is also a correct expansion for the acronym, which in this case serves as initials (as described in Section 3.7.4). Other names were frequently classified as expansions for various acronyms, likely because they have a high PMI. Of course, not all are necessarily "correct." For example, `aa"g` / **אא"ג** = `andrh agasi` / **אַנדרה אַגאסי** (*Andre Agassi*, "Andre Agassi (name)") is probably not a true acronym-expansion pair, as the formation rule is `[1,12]` instead of `[1,1]`,

which would be more likely for initials. (Note, however, that the acronym following the latter formation rule, a"a / א"א, *was* included by the classifier with the expansion, elsewhere in the dictionary.)

High PMIs and high frequencies likely account for most other probable misclassifications as well. For example, i"g / י"ג = iwtr grwyh / יותר גרועה (*more terrible*, "worse") and e"b / ש"ב = ebt birwelim / שבת בירושלים (*Sabbath in+Jerusalem*, "Sabbath in Jerusalem") are unlikely to be correct except in unusual cases, though the expansions have both high frequency and high PMI.

While in general it is difficult to state conclusively that an expansion is truly incorrect for an acronym, because it might simply be a rare usage, we can occasionally do so definitively. An example is the last example entry, plm"x / פלמ"ח = pnsiwni lmwrim xiilim / פנסיוני למורים חיילים (*pensions-related to+teachers soldiers*, "pensions for soldier-teachers"). This "expansion" is in fact part of a noun phrase, wtq pnsiwni lmwrim xiilim / ותק פנסיוני למורים חיילים (*seniority pension-related to+teachers soldiers*, "pension seniority for soldier-teachers"), which was a legislative issue discussed in a document from the Knesset corpus. We speculate that its very high PMI led to the classifier misclassifying it as an expansion for the acronym.

## 4.6 Extrinsic Evaluation: Acronym Disambiguation

An acronym in text can often have many possible expansions, only one of which is correct. *Disambiguation* is the process of determining which expansion is correct for the particular context. We addressed a variant of this problem as an extrinsic evaluation for the dictionary we built in Chapter 4. Its quality was assessed by evaluating the degree of improvement, compared to dictionaries built using existing methods, that it supplied to disambiguation efforts of acronyms in context.

### 4.6.1 Evaluation Set

A total of 202 acronym types in context were hand-analyzed, as previously described in Section 1.3.2. An example instance is the acronym btel"d / בתשל"ד in the following sentence from the HaAretz corpus:

hhitr hklli lybwdh beywt nwspwt nitn btel"d...
...ההיתר הכללי לעבודה בשעות נוספות ניתן בתשל"ד
*the+permission the+general for+work in+hours additional given [acronym]...*
"The general permission for additional working hours was given [acronym]..."

Human annotators provided the correct analysis for the acronym:

- *Prefix:* b+ / ב+ ("in / on").
- *Suffix:* None.
- *Expansion:* The Hebrew calendar year (5)734 (corresponding to the Gregorian calendar year 1973-74).[4]

A subset of 25 (12%) instances were reserved for development, and 10 were discarded as typos or errors, leaving 167 types. Of these, 25 were isopsephic acronyms.

The documents in which the instances appeared were of course held out of all procedures involved in the dictionary-building process described in Chapter 4, so as to be eligible for evaluation here. After the LDA model was trained on the other corpora documents (as described in Section 4.3.1), we inferred LDA topic scores for these held-out documents, which was useful for dictionary entry ranking.

### 4.6.2 Baselines

We compared the performance of the dictionary we built with two other dictionaries representing the existing state-of-the-art.

- *Baseline #1*: Inspired by the most common previous method of acronym dictionary-building (Section 2.1), we searched the corpora for Hebrew acronyms that were either immediately followed by a parenthetical clause of at least two words, or were themselves in parentheses and preceded by 2–4 words—for example, "CIA (Central Intelligence Agency)" or "Central Intelligence Agency (CIA)."

  Rather than re-implement existing algorithms, we manually annotated each such case as being a proper acronym / expansion match or

[4] See Section 3.7.2 for more on isopsephic acronyms.

not. We were generous in this assessment, even if the non-acronym part was not an *exact* match for the expansion: for example, the sentence fragment "CIA (the government officials at the Central Intelligence Agency)" was rated as providing a correct match, even though the parenthetical phrase contained extraneous words beyond the expansion.

This baseline thus served as an upper bound for the *best possible* acronym dictionary constructed from local parenthetical acronyms.

- *Baseline #2*: We used a simple combination of the three combined gold-standard dictionaries of human-curated acronym-expansion pairs (see Section 1.3.3). We augmented it with the same isopsephic entries we generated for the dictionary we built (see Section 4.4), as these were generally lacking from the gold-standard dictionaries yet trivial to add. We intentionally strengthened this baseline so as to demonstrate that any improvement in our dictionary's performance was due to our expansion-identification methods in corpora, not from the trivial process of generating isopsephic entries.

### 4.6.3 Dictionary Entry Ranking

For a given acronym, each dictionary typically had more than one—sometimes many—expansion possibilities. Therefore, ranking the expansions in the dictionary entry for a given acronym was an influential factor in performance on the disambiguation task.

First, for each acronym instance to disambiguate, we considered all possible prefix analyses if the acronym began with suitable letters. For example, `btel"d` / `בתשל"ד` could conceivably have been either a five-letter acronym or the four-letter acronym `tel"d` / `תשל"ד` prefixed with a `b+` / `ב+` ("in / on"). We considered the analyses in order from shortest to longest prefix; in this case, assuming there was the shortest possible prefix—none at all—and only afterwards guessing that the prefix was `b+` / `ב+` ("in / on"). This decision was based on the observation, from Table 3.6, that shorter prefixes are almost always more likely than longer ones; this inclination was proven beneficial when tested on the development set too.

For the dictionary we built, we ordered by source (as described in Section 4.4)—first isopsephic, then gold, then special, then classifier—as this

gave the best results on the development set. Then, we ranked entries by the LDA similarity score of the expansion and the acronym's document context.[5]

The entries for Baseline #2 (gold dictionary with isopsephic entries) had no natural ranking, so we ordered expansions at random within the entry, though again isopsephic expansions (if any) were always first. Typically there was only one or a few expansions per acronym entry in this dictionary, so the ranking was less important here.

We did not attempt ranking the entries for Baseline #1 (best-possible local parentheses dictionary) because, as we shall soon see, it performed very poorly even for the most generous case.

### 4.6.4 Results

We evaluated the three dictionaries—the one that we built, and the two baselines—with respect to the following test: Given an acronym and the document it appears in, is its correct expansion (with respect to its context) in the top $r$ results of the dictionary's entry for that acronym?

We tested four values of $r$: $r = 1$ ("is the very top expansion correct?"), $r = 2$ and $r = 3$ ("is the correct expansion in the top 2 (or 3) entries?"), and $r = \infty$ ("is the correct expansion in the dictionary at all for this acronym?"). Performance was measured as the percentage of instances of the evaluation set which passed this test.

Our dictionary performed well, and outperformed both baseline dictionaries, especially the first (best-possible dictionary of local parenthetical acronyms). Note that because of how we constructed and ranked the entries in our dictionary, it is guaranteed to perform at least as well as the strong Baseline #2; what we are interested in is how *much* better. Since Baseline #2 had very high performance as well, looking at the error rate reduction of our dictionary was a better measure of improvement. The $p$ values were calculated using McNemar's paired $\chi^2$ one-tailed test, and show statistically-significant improvement for the most important $r = 1$ case, and possibly-significant improvement for the $r = 3$ and $r = \infty$ cases (using the conventional 0.05 significance level threshold).

[5]The calculation was identical to that described in Section 4.3.1, where we computed the LDA similarity score of the acronym and possible expansion $n$-gram. Here, we replaced the acronym topic vector $\overrightarrow{a}$ with the inferred document topic vector.

| Dictionary | $r = 1$ | $r = 2$ | $r = 3$ | $r = \infty$ |
|---|---|---|---|---|
| *Baseline #1*: Best-possible dictionary of local parenthetical acronyms | | | | 52.38% |
| *Baseline #2*: Gold dictionaries with isopsephic entries | 66.47% | 77.25% | 78.44% | 82.63% |
| Our dictionary | 72.46% | 79.04% | 81.44% | 85.03% |
| Error rate reduction (%) of our dictionary vs. Baseline #2 | 17.86 ($p < 0.03$) | 7.89 ($p < 0.25$) | 13.89 ($p < 0.06$) | 13.79 ($p < 0.06$) |

Table 4.6: Performance of the dictionaries on the disambiguation task, given as the percentage of the evaluation set instances which have the correct expansions in the top $r$ results for the dictionary's entries for the acronym.

### 4.6.5 Error Analysis

Disambiguation errors fell into two broad categories:

1. Acronyms whose correct expansions (for the given context) did not appear at all in the acronym's dictionary entry; and
2. Acronyms whose correct expansions did appear, but ranked below other incorrect expansions (for the given context, or at all).

The two types can be distinguished by comparing the performance on the $r = \infty$ case to the finite-$r$ cases.

For errors of the first type, the correct expansion was generally too rare to have been caught by the dictionary-building algorithm, or to be well-known enough to have a correct expansion in the gold dictionary. Sometimes these acronyms were local acronyms, with the expansion appearing nearby, generally for rare or document-specific acronyms. This error class points to the future utility of combining acronym dictionary-building techniques—both ours and the more "traditional" methods of searching for expansions in the text surrounding the acronym, especially in parentheses. (The latter technique *alone*, however, performs very poorly, as shown by the lackluster performance of Baseline #1.)

Acronyms of the second error class—where the correct expansion *was* in the dictionary entry, just ranked after other, incorrect expansions—comprised about 13% of all instances, and 46% of the instances which did not have the correct expansion as the acronym's very first entry.

As described in Section 4.6.3, our dictionary's entries were ranked by the source of the expansion: first isopsephic, then gold, then special, then classifier. Most acronyms did not have expansions of more than one or two sources (in fact, not a single one of the evaluation set acronyms had a special-source expansion). As shown by the high performance of Baseline #2 in Table 4.6, most acronyms' correct expansions were isopsephic and/or gold-sourced.

Sometimes there were multiple gold expansions in our dictionary which did not appear in the corpora frequently enough to have LDA score values. They were then ranked at random amongst the top entries, and therefore the correct expansion (whether gold-sourced or otherwise) could be pushed further down the list. For example, in one instance the acronym `rbe"y` / **רבש"ע** in context meant `ribwnw el ywlm` / **ריבונו של עולם** (*master-of+him of world*, "Master of the World / God"). The top three dictionary entries for the acronym were:

1. an incorrect expansion, from the gold source;
2. the correct expansion, from the gold source; and
3. the correct expansion, with a variant spelling that prevented it from being combined with the similar gold-source expansion above, from the classifier source.

Despite errors of this type, ranking the gold entries first was still a sound strategy, and doing otherwise would have resulted in lower performance overall as determined by experiments on the held-out development set.

By construction, all isopsephic acronyms (which, recall from Section 3.7.2, comprise a hefty 16% of acronym types) were correctly expanded by their very first entries in both our dictionary and the second baseline.

# Chapter 5

# Discussion

## 5.1 Conclusions

We developed a new machine learning method to automatically create an acronym dictionary from unstructured corpora. Unlike prior methods, this approach includes global acronyms (those unaccompanied by their expansions in the same document). Dictionaries built with our method are easily updatable as new acronyms are invented, can be applied to specialized genres, and are more comprehensive than human-curated dictionaries. Additionally, contextual data is included about the expansions that helps determine which of the (possibly multiple) dictionary expansions is the correct one for a given acronym instance in text.

Our method was applied to Hebrew corpora to create a new Hebrew language resource, useful for natural language processing tasks.

As a means of extrinsically evaluating the dictionary's quality, we applied it to the problem of acronym disambiguation in context on 167 instances. We succeeded in identifying the correct expansion in 72.46% of the instances, achieving the statistically-significant error rate reduction of 17.86% over a strong baseline (a human-edited acronym dictionary enhanced with generated isopsephic entries). Additionally, the correct expansion was included for the acronym in our dictionary—albeit ranked after other expansions which were incorrect for the context—in 85.03% of the acronym instances, an error rate reduction of 13.79% compared to the baseline. We also compared our dictionary's performance to that of another baseline dictionary (constructed from local parenthetical acronyms, the leading prior technique for automatic

acronym dictionary-building) and achieved an error rate reduction of 69%.

## 5.2 Future Work

### 5.2.1 Specialized Hebrew Domains

Our work focused on secular Modern Hebrew texts, but our methods are easily adaptable to specialized genres of text by substituting corpora and a genre-specific gold-standard dictionary. Obvious areas include:

- **Military texts**, which are a especially rife with acronym usage. The Israel Defense Forces published a large dictionary of military acronyms [11] that would be useful as a training and evaluation set.
- **Jewish legal texts**, which already have comprehensive human-edited dictionaries for their mostly-closed set of acronyms. It would be interesting to compare performance on the task of disambiguating Jewish legal acronyms in texts with the methods of HaCohen-Kerner et al. [18]. They used machine learning techniques starting with the existing resource of an acronym-expansion dictionary, unlike our methods which generate such a dictionary, including contextual topic information.

Another possibility is to apply the classifier we developed in Chapter 4—which was trained on general Modern Hebrew—to domains without existing gold-standard dictionaries (such as internal corporate documents, specialized research fields, etc.). Note however that not all domains that are common foci of acronym research in other languages, like the biomedical research field extensively studied in English, are suitable for Hebrew acronyms because technical terms are usually not actually written in Hebrew, or tend to be transliterated or translated (see Section 3.7.1).

### 5.2.2 Other Languages

Every language has its own set of challenges and opportunities for acronym identification, dictionary-building, and disambiguation. The main advantages of Hebrew are the ease of identifying acronyms due to their specialized orthographic styling (as described in Section 3.1), and the widespread usage

of acronyms in texts. The main disadvantages are the relatively modest array of natural language processing resources in Hebrew; and the complexity introduced by Hebrew's complicated morphology and orthography, particularly prefixed function words (detailed in Section 3.5).

Other languages with complicated morphologies (like Arabic) and/or constrained by poor levels of language-processing resources (like most non-English languages) may especially benefit from our approaches. For languages with non-trivial acronym identification, including English and Arabic, our work would need to be combined with more sophisticated methods of identifying acronyms.

### 5.2.3 Named Entity Recognition and Multi-Word Expressions

A classic natural language processing challenge is *named entity recognition*, the identification of people or organizational names, locations, geographical locations, times, etc. This task is especially difficult in languages like Hebrew [25] and Arabic [44], which lack the capitalized letters that are strong clues in other languages like English.

Another common problem is identifying and interpreting *multi-word expressions*, phrases with idiosyncratic meanings not predictable from the individual words [42]. For example, the English multi-word expression "kick the bucket" means "die" rather than "hit the bucket with a foot."

We observed that many acronym expansions are named entities, like the Israeli city q"e / ק"ש = q̲rit e̲mwnh / קְרִית שְמונה (*city-of eight*, "Kiryat Shmonah"); or multi-word expressions, such as ayp"k / אעפ"כ = a̲p y̲l p̲i k̲n / אַף עַל פִּי כֵן (*even on by thus*, "even though"). (Of course, the converse does not hold: most named entities and multi-word expressions are not acronyms.) Thus, we suggest that using an acronym dictionary will strengthen investigations of understanding and recognizing named entities and multi-word expressions.

### 5.2.4 Additional Extrinsic Evaluations

We chose a natural extrinsic evaluation task—acronym disambiguation in context—to capitalize on the novel context-related metadata of the entries of the acronym dictionary we build. Other intuitive evaluation possibilities

include assessing use of the acronym dictionary's impact on information retrieval or machine translation. We leave these directions for future research.

# Appendix A

# Latent Dirichlet Allocation (LDA) Topic Models

Topic modeling algorithms are methods for discovering hidden (latent) themes in large textual datasets. One commonly-used model is Latent Dirichlet Allocation (LDA), developed in 2003 by David Blei, Andrew Ng and Michael Jordan [4]. We applied an LDA model's information during dictionary-building to help match acronyms and their expansions, using the assumption that they will share similar topic proportions (see Section 4.3.1); and later to aid acronym disambiguation in context, by choosing the acronym's expansion that is closest to the context in the topic space (see Section 4.6.3).

LDA represents each document as a probability distribution over a set of topics; each topic is in turn a probability distribution over a set of words. LDA uses a "bag of words" assumption which ignores word order within documents. To reduce noise in the model, we removed all stop words during pre-processing.

The LDA topic modeler takes as input a collection $D$ of textual documents, and a parameter $T$ fixing the total number of topics to discover in those documents. It returns as output:

- a list of topics, each represented by a short ranked list of words most associated with it (note that word types can appear in more than one topic);

- for each document, the proportion of the document belonging to each topic (note not all words in a topic need appear in the document for it to be associated); and
- for each word type, the proportion of the word that is associated with each topic.

Consider the simplified illustrative example in Figure 5.1. The topic model analyzed a collection of documents and discovered three topics: topic #1 contains healthcare-related words, topic #2 contains finance-related words, and topic #3 contains cat-related words.[1] A particular document on hospital funding cuts contains several words from topic #1, a few from topic #2, and none from topic #3; it is thus represented as a weighted combination of the topics with more weight to topic #1 and none to topic #3.

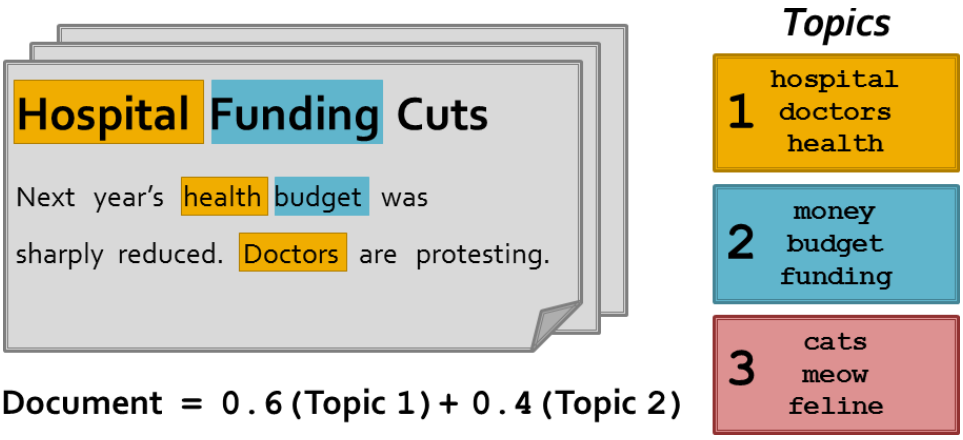



Figure 5.1: A simplified illustrative example of an LDA topic model. Three topics (sets of associated words) are "discovered" in a collection of documents. Each document is then represented as a weighted combination of topics.

Since the LDA topic model requires the parameter $T$ (the number of topics for the model) to be fixed in advance, determining the "right" value

[1] In practice, there are more words per topic, and relevancy scores associated with each word. Additionally, while these topics are easily intuitively "labeled" as relating to healthcare, finance, and cats, in actuality it is not always clear what human-recognizable concept is represented by a topic; see [29].

for $T$ is not trivial. Intuitively, too small a $T$ results in overly coarse topics; too large a $T$ results in overly fine and statistically-meaningless topics. For the LDA features we calculated in Section 4.3.1, we tried a range of values for $T$ (between 200 and 600 in increments of 50) and simply selected the value that gave the best classification results, $T = 300$.

While the LDA model is fixed once it is built, topic scores can easily be inferred for new documents that were not included in $D$ when the model was being built. We did this for the held-out documents of the disambiguation task's evaluation set (described in Section 4.6.1).

For further background on LDA, see [4] or [3].

# Appendix B

# Whimsy

During our acronym research, we encountered much wonderful whimsy.

- Author Douglas Adams quipped: "The World Wide Web is the only thing I know of whose shortened form takes three times longer to say than what it's short for."[1]
- A delightful meta-acronym entry in the online Hebrew dictionary Wikimilon [51] is `rtl"m` / `רתל"ם` = `r̲aei t̲ibwt l̲a m̲wbnim` / **רַאשֵׁי תֵּיבוֹת לֹא מְובָנִים** (*heads-of letters not understood*, "acronyms that are not understood").
- The South Lake Union Trolley, a public transit system in Seattle, WA, was hastily renamed the South Lake Union Streetcar after officials realized the ill-advised nature of its original name's acronym.[2]
- German favors creating acronyms from syllables rather than letters, such as Gestapo instead of GSP for Geheime Staatspolizei (*secret state-police*, "Secret State Police"). Sometimes this can reach comic extents, as in Vokuhila = vorne kurz, hinten lang (*short in the front, long in the back*, "mullet"), a practice referred to by the acronym AbKüFi = Abkürzfimmel ("strange habit of abbreviating").

[1] *The Independent on Sunday*, 1999.

[2] S. Pomper. *Seattle Curiosities: Quirky Characters, Roadside Oddities, & Other Off-beat Stuff.* Globe Pequot, 2009.

- The world's longest acronym is a 54-letter Cyrillic acronym in the 1969 *Concise Dictionary of Soviet Terminology*, with the English-translated expansion "The laboratory for shuttering, reinforcement, concrete and ferroconcrete operations for composite-monolithic and monolithic constructions of the Department of the Technology of Building-assembly operations of the Scientific Research Institute of the Organization for building mechanization and technical aid of the Academy of Building and Architecture of the Union of Soviet Socialist Republics."[3]

- The longest Hebrew acronym in the corpora was a satire of the long, laudatory Jewish religious acronym style, used with a non-Jewish religious meaning meant as an insult: zcwqlyemrwlh"h / זצוקלעשמרולה"ה = zkr cdiq wqdwe lbrkh yl emw mnzrim rbim wlxii hywlm hba / זכר צדיק וקדוש לברכה על שמו מנזרים רבים ולחיי העולם הבא (*in-memory-of righteous-person and+sanctify to+blessing on name+his monastaries many and+to+life the+world the+next*, "in memory of a righteous person and may his name be sanctified with blessings on many monastaries for the eternal afterlife").

- Acronyms can cross language boundaries in Jewish tradition. Rabbi Meir of 19th century Farmishlan, Poland interpreted Hebrew acronyms derived from Biblical passages by their "expansions" in Polish.[4]

- The 16th century Egyptian Jewish community leader, Rabbi David son of Shlomo son of Zimrah, recorded a story in his responsa 2322: "It happened that someone brought to me a contract which had written on the bottom 'And on this nq"s / נק"ש,' and I couldn't understand what it meant, until one of the litigants came to me and explained that it was an acronym for nvl qnin elm / נטל קנין שלם (*received purchase complete*, "the purchase was completely received"). I said to him, 'You need to walk around with this contract everywhere it goes in order to explain it, because I think it should be an acronym for nqra qvn ewvh / נקרא קטן שוטה (*he-is-called little fool*, "he's called a little fool")."'

[3] A. Cantrell. *The Book of Word Records: A Look at Some of the Strangest, Shortest, Longest, and Overall Most Remarkable Words in the English Language*. Adams Media, 2013.

[4] Y. Spiegel. כוחם של ראשי תיבות וגימטריות (Power of Acronyms and Gematrias). ירושתנו *(Our Inheritance)*, (9), 2012. In Hebrew.

- Hebrew acronym misunderstandings have a long history of affecting Jewish religious customs and legal decisions. A famous acronym mistake is seen towards the end of the Grace After Meals liturgy, which quotes from the Biblical verse of Psalms 18:51:

  `mgdl iewywt mlkw wyeh xsd lmeixw ldwd wlzryw yd ywlm.`
  **מגדל ישועות מלכו ועשה חסד למשיחו לדוד ולזרעו עד עולם.**
  *magnifies salvation king+his and+does mercy to+annointed+his to+David and+to+descendants+his until ever.*
  "He magnifies salvation for His king and deals kindly with His anointed, with David and his descendants forever."

  On the Jewish Sabbath and holidays, however, Grace-reciters instead substitute a nearly-identical verse from II Samuel 22:51. The first word is replaced by `mgdwl` / **מגדול** (*tower*, "tower"), changing the verse's meaning to "He is a tower of salvation for His king..." Why?

  According to the prominent Rabbi Baruch Halevi Epstein in his 1940 book, **ברוך שאמר** (Praised is He Who Speaks), the Psalms version is the correct one for all occasions, including the Sabbath and holidays. However, Epstein explains, a long-ago printer of the Grace After Meals drew attention to the similar verse in II Samuel through a margin note of `bemwal b': mgdwl` / **בשמואל ב': מגדול** (*in+Samuel II: tower*, "in II Samuel: tower"). In a subsequent printing, it was shortened to the acronym `be"b: mgdwl` / **בש"ב: מגדול**. Later, this acronym was misunderstood to mean `bebt: mgdwl` / **בשבת: מגדול** (*in+Sabbath: tower*, "on the Sabbath: tower"), and written out in long-form. Finally, people reasoned that a liturgical change on the Sabbath was surely appropriate for holidays as well. Thus developed the convention of reading the II Samuel version on the Sabbath and on Jewish holidays, and the original Psalms version otherwise. And all from an acronym misunderstanding!

  Sadly, the historical accuracy of this delightful story has recently been contested,[5] but it still serves as an excellent method of piquing Hebrew acronym interest in religious Jews.

[5] R. Apple. Magdil and migdol—Liturgical responses to textual variants. *The Jewish Bible Quarterly*, 41, April–June 2013.

- Recursive acronyms refer to *themselves* in their expansion, sometimes humorously. These were especially common in the early computer hacker community. For example, the Unix-like computer operating system GNU stands for GNU's Not Unix.[6]
- Nested acronyms can occur without recursion, too. The New Scientist magazine ran an informal competition for the deepest-nested acronym example,[7] and the winner was RARS = regional ATOVS retransmission service, which included ATOVS = advanced TOVS, which included TOVS = TIROS operational vertical sounder, which included TIROS = television infrared observational satellite, for a total of four acronym levels.
- In 2009, NASA ran a public contest to name a module of the International Space Station. Television comedian Stephen Colbert encouraged his show's viewers to vote to name it after him, and won by over 40,000 votes. While NASA eventually chose the more traditional second-place name instead, it offered consolation by naming the station's exercise equipment the Combined Operational Load-Bearing External Resistance Treadmill.[8]
- The U.S. defense agency DARPA has a long tradition of humorously naming projects with contrived acronyms.[9,10] For example, NACHOS = Nanoscaled Architecture for Coherent Hyper-Optic Sources, BATMAN = Biochronicity and Temporal Mechanisms Arising in Nature, and ROBIN = Robustness of Biologically-Inspired Networks.

[6] R. Stallman. "The Free Software Movement and the Future of Freedom." (Speech, Zagreb, Croatia, March 9, 2006). *Free Software Foundation Europe.* `http://fsfe.org/freesoftware/transcripts/rms-fs-2006-03-09.en.html#the-name-gnu`. Accessed February 3, 2014.

[7] Very deeply nested acronyms. *New Scientist*, 2768, July 7, 2010.

[8] S. Siceloff. COLBERT ready for serious exercise. *National Aeronautics and Space Administration*, May 5, 2009. `http://www.nasa.gov/mission_pages/station/behindscenes/colberttreadmill.html`. Accessed February 3, 2014.

[9] N. Hodge. "Darpa's nanoscale NACHOS and other awesome acronyms. *Wired.com: Danger Room*, May 22, 2009. `http://www.wired.com/dangerroom/2009/05/darpas-nanoscale-nachos`. Accessed February 3, 2014.

[10] M. Hardy. Batman and Robin's new secret hideout: DARPA. *GCN*, July 8, 2010. `http://gcn.com/Articles/2010/07/08/batman-robin-darpa-acronyms.aspx`. Accessed February 3, 2014.

- Nineteenth century author Edgar Allan Poe was one of the original explorers of the intersection between acronyms and whimsy. Consider the following excerpt from his satirical short story, *How to Write a Blackwood Article*:[11] "[...] [E]verybody has heard of me. I am [...] so justly celebrated as corresponding secretary to the 'Philadelphia, Regular, Exchange, Tea, Total, Young, Belles, Lettres, Universal, Experimental, Bibliographical, Association, To, Civilize, Humanity.' Dr. Moneypenny made the title for us, and says he chose it because it sounded big like an empty rum-puncheon. (A vulgar man that sometimes—but he's deep.) We all sign the initials of the society after our names, in the fashion of the R. S. A., Royal Society of Arts—the S. D. U. K., Society for the Diffusion of Useful Knowledge, &c, &c. Dr. Moneypenny says that S. stands for stale, and that D. U. K. spells duck, (but it don't), that S. D. U. K. stands for Stale Duck and not for Lord Brougham's society—but then Dr. Moneypenny is such a queer man that I am never sure when he is telling me the truth. At any rate we always add to our names the initials P. R. E. T. T. Y. B. L. U. E. B. A. T. C. H.—that is to say, Philadelphia, Regular, Exchange, Tea, Total, Young, Belles, Lettres, Universal, Experimental, Bibliographical, Association, To, Civilize, Humanity—one letter for each word, which is a decided improvement upon Lord Brougham. Dr. Moneypenny will have it that our initials give our true character—but for my life I can't see what he means."

- The role-playing game Dungeons and Dragons includes a player role called the Dungeon Master Helper, referred to by its acronym DMH, which is pronounced "dee em aitch." The Hebrew version is not the expected script transliteration `dm"h` / `דמ"ה` but rather the *phonetic* transliteration `dm"c'` / `דמ"צ'`, pronounced "da match."[12]

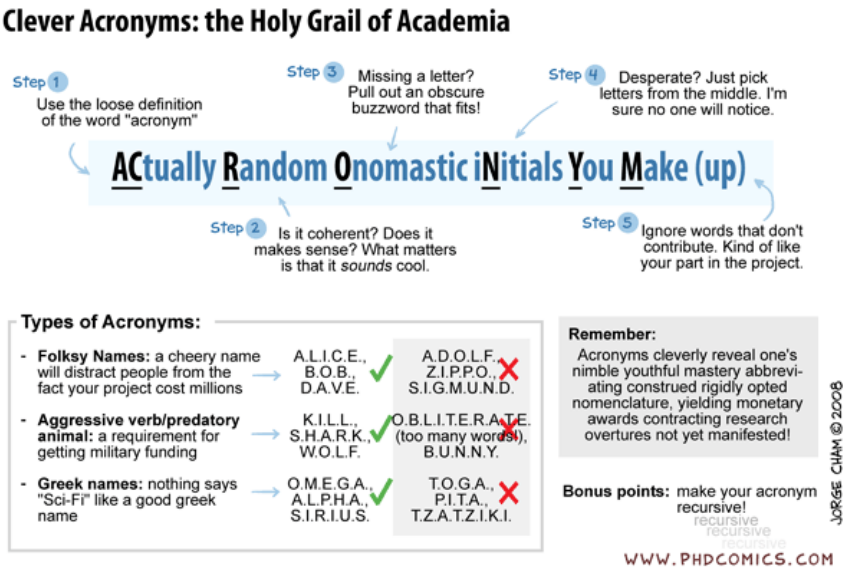
- The collectible trading card game *Magic: The Gathering* has a card titled "Our Market Research Shows That Players Like Really Long Card Names So We Made this Card to Have the Absolute Longest Card Name Ever Elemental," with the description "Just call it OMRSTPLRLCNSWMTCTHTALCNEE for short."[13]

[11] E. Poe. *How to Write a Blackwood Article.* 1850.

[12] Thanks to Tomer Wintner for bringing this to our attention.

[13] Wizards of the Coast. Our Market Research Shows That Players Like Really Long

- An English acronym following an extremely unusual formation rule is the mathematical and philosophical term IFF = if and only if, which was likely constructed thus for its similarity to the word "if." The Hebrew equivalent, `am"m` / `אם"ם` = `am wrq am` / `אם ורק אם` (*if and+only if*, "if and only if") shares this rare formation rule, with an exceptional appearance of a final-form Hebrew letter (`m` / `ם` as opposed to the regular `m` / `מ`) in the *middle*, not end, of the word.[14]
- Acronyms can sometimes be part of multi-word expressions that contain redundant words, a phenomenon known as RAS Syndrome, short for the wonderfully-named Redundant Acronym Syndrome Syndrome. For example, ATM machine = automated teller machine machine or `pyilwt px"y` / `פעילות פח"ע` = `pyilwt pyilwt xblnit ywint` / `פעילות פעילות חבלנית עוינת` (*operations operations damaging hostile*, "terrorism attack").




*Credit: "Piled Higher and Deeper" by Jorge Cham,* `www.phdcomics.com`.

Card Names So We Made this Card to Have the Absolute Longest Card Name Ever Elemental. *Magic: The Gathering Gatherer.* `http://gatherer.wizards.com/pages/Card/Details.aspx?multiverseid=74237`. Accessed February 3, 2014.

[14]Thanks to Nachum Dershowitz for bringing this to our attention.

**Thesis in limerick form:**

Some phrases or names can be long,
And the urge to shorten 'em strong,
    No need feeling queasy—
    Compressing is easy...
Though pause lest you do it all wrong!

Take from each word the first letter
(Although, two or more can be better),
    Then aggregate them,
    Concatenate them,
And an acronym you'll getter.

But read acronyms inside text,
And you might stop abruptly, quite vex'd,
    Its meaning, see, might
    Not quite come to light,
Thus leaving you very perplex'd.

For acronym meanings are rife,
Depending on which part of life
    They happen t'address—
    Forcing you to guess,
And causing confusion and strife.

Dictionaries address this fright,
And dictionaries people do write.
    But they're expensive,
    Not comprehensive,
And don't know which meaning is right.

So an acronym dictionary *we* build:
From text automatically distill'd.
    A computer reads
    (At very high speeds)
And magic'ly, entries are filled.

Contextual info we include,
So that when the acro's then view'd,
    You don't feel harried
    By answers varied,
And the best match you can conclude.

On Hebrew we use our technique,
Which gives opportunity t'seek
    Insights linguistic,
    Moreover statistic,
On acronyms' Hebrew mystique.


"Oh, it's an acronym for 'It doesn't stand for anything.'"
*Credit: Harley Schwadron.*

# Bibliography


[1] S. Ashkenazi, D. Jarden, and J. Stevens. **אוצר ראשי תיבות** *(Treasury of Acronyms)*. Kiryat Sefer, Jerusalem, 1994. In Hebrew.

[2] O. Bat-El. The optimal acronym word in Hebrew. In Koskinen, editor, *Toronto Working Papers in Linguistics*, pages 23–37. Toronto Working Papers in Linguistics, 1994.

[3] D. Blei. Probabilistic topic models. *Communications of the ACM*, 55(4):77–84, 2012.

[4] D. Blei, A. Ng, and M. Jordan. Latent Dirichlet Allocation. *Journal of Machine Learning Research*, 3:993–1022, 2003.

[5] S. Bolozky. On the special status of the vowels a and e in Israeli Hebrew. *Hebrew Studies*, 40:233–250, 1999.

[6] C. Chang and C. Lin. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2:27:1–27:27, 2011.

[7] D. Crystal. *Language and the Internet*. Cambridge University Press, 2nd edition, 2001.

[8] D. Dannélls. Acronym recognition: Recognizing acronyms in Swedish texts. Master's thesis, Department of Linguistics, University of Gothenburg, Gothenburg, Sweden, June 2006.

[9] D. Dannélls. Automatic acronym recognition. In *Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics*, pages 167–170, Trento, Italy, 2006.

[10] D. Dannélls. Acronym classification using feature combinations, 2007.

[11] Israel Defense Forces. **מילון הקיצורים וראשי התיבות** (Dictionary of abbreviations and acronyms), 2010. In Hebrew.

[12] G. Fu, K. Luke, G. Zhou, and R. Xu. Automatic expansion of abbreviations in Chinese news text. In H. Tou Ng, editor, *Information Retrieval Technology: 3rd Asia Information Retrieval Symposium (AIRS 2006)*, Singapore, October 2006.

[13] S. Gaudan, H. Kirsch, and D. Rebholz-Schuhmann. Resolving abbreviations to their senses in MEDLINE. *Bioinformatics*, 21:3658–3664, 2005.

[14] L. Glinert. *The Grammar of Modern Hebrew*. Cambridge University Press, 1989.

[15] Y. HaCohen-Kerner, A. Kass, and A. Peretz. Baseline methods for automatic disambiguation of abbreviations in Jewish law documents. In *Proceedings of the 4th International Conference on Advances in Natural Language*, 2004.

[16] Y. HaCohen-Kerner, A. Kass, and A. Peretz. Abbreviation disambiguation: Experiments with various variants of the one sense per discourse hypothesis. In *Proceedings of the Application of Natural Language to Information Systems (NLDB'08)*, pages 27–39, 2008.

[17] Y. HaCohen-Kerner, A. Kass, and A. Peretz. Combined one sense disambiguation of abbreviations. In *ACL (Short Papers)*, pages 61–64, 2008.

[18] Y. HaCohen-Kerner, A. Kass, and A. Peretz. HAADS: A Hebrew Aramaic abbreviation disambiguation system. *Journal of the American Society for Information Science and Technology*, 61(9):1923–1932, 2010.

[19] Y. HaCohen-Kerner, A. Kass, and A. Peretz. Initialism disambiguation: Man versus machine. *Journal of the American Society for Information Science and Technology*, 64(10):2133–2148, 2013.

[20] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. Witten. The WEKA data mining software: an update. *SIGKDD Explorations*, 11(1):10–18, November 2009.

[21] A. Itai and S. Wintner. Language resources for Hebrew. *Language Resources and Evaluation*, 42(1):75–98, 2008.

[22] A. Jain, S. Cucerzan, and S. Azzam. Acronym-expansion recognition and ranking on the web. In *Information Reuse and Integration (IRI 2007)*, pages 209–214. IEEE, August 2007.

[23] X. Ji, G. Xu, J. Bailey, and H. Li. Mining, ranking, and using acronym patterns. In *Proceedings of the 10th Asia-Pacific Web Conference on Progress in WWW Research and Development (APWeb'08)*, pages 371–382, Berlin, Heidelberg, 2008. Springer-Verlag.

[24] Kizur.co.il. **מילון קיצורים וראשי תיבות** (Dictionary of abbreviations and acronyms). In Hebrew.

[25] G. Lemberski. Named entity recognition in Hebrew language; Hebrew multiword expression: approaches and recognition methods. Master's thesis, Ben-Gurion University, 2003.

[26] C. Mair. *Twentieth-Century English: History, Variation and Standardization*. Studies in English Language. Cambridge University Press, 2009.

[27] L. Marwick. *Biblical and Judaic Acronyms*. KTAV Publishing House, Brooklyn, NY, 1979.

[28] A. McCallum. MALLET: A machine learning for language toolkit. http://mallet.cs.umass.edu, 2002.

[29] Q. Mei, X. Shen, and C. Zhai. Automatic labeling of multinomial topic models. In Pavel Berkhin, Rich Caruana, and Xindong Wu, editors, *Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 490–499, August 2007.

[30] I. Meir. Morphological levels and diachronic change in modern Hebrew plural formation. *Studies in Language*, 30(4):756–827, 2006.

[31] Melingo. **מילון מורפיקס** (Morfix dictionary). `http://milon.morfix.co.il`. In Hebrew.

[32] M. Muchnik. **היביטים מורפו-פונמיים של הנוטריקון בעברית בת-ימינו** (Morpho-phonemic characteristics of acronyms in contemporary Hebrew). *Hebrew Linguistics*, 54:53–66, 2004. In Hebrew.

[33] D. Nadeau and P. Turney. A supervised learning approach to acronym identification. In *8th Canadian Conference on Artificial Intelligence (AI2005)*, pages 319–329, 2005.

[34] The Academy of the Hebrew Language. **גרשיים** (") (Double-quote mark (")). `http://hebrew-academy.huji.ac.il/hahlatot/Punctuation/Pages/P33.aspx`. In Hebrew. Accessed August 8, 2010.

[35] The Academy of the Hebrew Language. **תפקיד או תואר האישה** (Role or degree of the woman). `http://hebrew-academy.huji.ac.il/sheelot_teshuvot/Pages/02061001.aspx`. In Hebrew. Accessed August 8, 2010.

[36] K. Papineni. Why inverse document frequency? In *Proceedings of the second meeting of the North American Chapter of the Association for Computational Linguistics on Language technologies (NAACL '01)*, pages 1–8, Stroudsburg, PA, USA, 2001. Association for Computational Linguistics.

[37] Y. Park and R. Byrd. Hybrid text mining for finding abbreviations and their definitions. In *Proceedings of the 2001 Conference on Empirical Methods in Natural Language Processing*, pages 126–133, 2001.

[38] J. Platt. Fast training of support vector machines using sequential minimal optimization. In B. Schoelkopf, C. Burges, and A. Smola, editors, *Advances in Kernel Methods - Support Vector Learning*. MIT Press, 1998.

[39] J. Pustejovsky, J. Castano, B. Cochran, M. Kotecki, and M. Morrell. Automatic extraction of acronym-meaning pairs from MEDLINE databases. *Medinfo*, 10:371–375, 2001.

[40] R. Quinlan. *C4.5: Programs for Machine Learning*. Morgan Kaufmann Publishers, San Mateo, CA, 1993.

[41] D. Ravid. Internal structure constraints on new-word formation devices in modern Hebrew. *Folia Linguistica*, 24:289–348, 1990.

[42] I. Sag, T. Baldwin, F. Bond, A. Copestake, and D. Flickinger. Multiword expressions: A pain in the neck for NLP. In *Proceedings of the 3rd International Conference on Intelligent Text Processing and Computational Linguistics (CICLing-2002)*, pages 1–15, 2001.

[43] A. Schwartz and M. Hearst. A simple algorithm for identifying abbreviation definitions in biomedical texts. In *Proceedings of the Pacific Symposium on Biocomputing*, pages 451–462, 2003.

[44] K. Shaalan and H. Raza. NERA: Named entity recognition for Arabic. *Journal of the American Society for Information Science and Technology*, 60(8):1652–1663, 2009.

[45] K. Sima'an, A. Itai, Y. Winter, A. Altman, and N. Native. Building a tree-bank of Modern Hebrew text. *Traitment Automatique des Langues*, 42(2), 2001.

[46] Y. Spiegel. **השימוש בקיצורים ובראשי תיבות שאינם שכיחים** (The use of uncommon abbreviations and acronyms). **ישורון** *(The People of Israel)*, 2002. In Hebrew.

[47] Y. Spiegel. **עמודים בתולדות הספר העברי: הגהות ומגיהים** *(Pages in the History of the Hebrew Book: Glosses and Proof-readers)*. Bar-Ilan University Press, January 2005. In Hebrew.

[48] Y. Spiegel. **כוחם של ראשי תיבות וגימטריות** (Power of acronyms and gematrias). **ירושתנו** *(Our Inheritance)*, (9), 2012. In Hebrew.

[49] A. Stiensaltz. **ראשי תיבות וקיצורים בספרות החסידות והקבלה** *(Acronyms and Abbreviations in Hasidism and Kabbalah)*. Sifriyati, Tel Aviv, 1968. In Hebrew.

[50] U. Tadmor. **הנוטריקון בעברית הישראלית** (The acronym in Israeli Hebrew). **לשוננו לעם** *(Our Language for the People)*, 39:225–257, 1988. In Hebrew.

[51] Wiktionary. **ויקימילון** (Wikimilon). `http://he.wiktionary.org`. In Hebrew.

[52] J. Xu and Y. Huang. Using SVM to extract acronyms from text. *Soft Computing - A Fusion of Foundations, Methodologies and Applications*, 11:369–373, November 2006.

[53] J. Yi and N. Sundaresan. Mining the web for acronyms using the duality of patterns and relations. In *Proceedings of the 2nd International Workshop on Web Information and Data Management*, WIDM '99, pages 48–52, New York, NY, USA, 1999. ACM.

[54] G. Zadok. Abbreviations: A unified analysis of acronym words, clippings, clipped compounds, and hypocoristics. Master's thesis, Tel Aviv University, 2002.

[55] M. Zahariev. Efficient acronym-expansion matching for automatic acronym acquisition. In *Proceedings of the International Conference on Information and Knowledge Engineering*, pages 32–37, 2003.

# ראשי תיבות בעברית:
# זיהוי, פיענוח והתרת רב משמעות

קיילה ג'קובס

# ראשי תיבות בעברית:

# זיהוי, פיענוח והתרת רב משמעות

חיבור על מחקר

לשם מילוי חלקי של הדרישות לקבלת התואר

מגיסטר למדעים במדעי המחשב

**קיילה ג'קובס**

הוגש לסנט הטכניון – מכון טכנולוגי לישראל

תשרי ה'תשע"ה חיפה אוקטובר 2014

המחקר נעשה בהנחיית פרופ׳ אלון איתי בפקולטה למדעי המחשב בטכניון ופרופ׳ שולי וינטנר בחוג למדעי המחשב באוניברסיטת חיפה.

אני מודה לטכניון על התמיכה הכספית הנדיבה בהשתלמותי.

# תקציר


ראשי תיבות הם מילה שנוצרת בדרך כלל מהאותיות התחיליות של שתי מילים או יותר, שנקראות המשמעות של ראשי התיבות. לדוגמה, **בי"ד** הוא ראשי תיבות שלהן המשמעות **בֵּית דִּין**, אם כי משמעותן יכולה להיות גם **ב־14**.

לעברית יש היסטוריה ארוכה של שימוש בראשי תיבות, החל מתקופת המשנה. ראשי תיבות נפוצים במיוחד בסוגות הייעודיות של מסמכים הקשורים ליהדות והלכה בתקופות השונות, כמו גם בטקסטים צבאיים ישראליים. ככלל, בטקסטים החילוניים בעברית מודרנית, אותם חקרנו, ראשי תיבות מהווים כ־1% מכל תמניות המלים, וכ־3% מתבניות המלים. ראשי תיבות בעברית נחקרו בעבר מנקודת מבט בלשנית, אך מעולם לא מזווית כמותית או סטטיסטית.

זיהוי של ראשי תיבות והבנת משמעותם יכולים לסייע במגוון רחב של יישומים חישוביים, כגון:

- **שליפת מידע:** בחיפוש אחר מסמך באמצעות שאילתא המכילה ראשי תיבות, יש להחזיר גם מסמכים שמכילים את משמעות ראשי התיבות, ולהפך.
- **תרגום מכונה:** בתרגום אוטומטי של טקסט משפה אחת לאחרת, פעמים רבות ראשי תיבות מהווים אתגר. אם טקסט המקור מכיל ראשי תיבות, אין זה מספיק, בדרך־כלל, רק לתעתק את האותיות של ראשי התיבות; ייתכן וראשי התיבות כלל אינם קיימים בשתי השפות.
- **הבנה או התרת רב־משמעות של מובנם של ראשי תיבות:** ייתכן וראשי תיבות בטקסט אינם מוכרים לקורא–בין אם הקורא אנושי או ממוחשב–וכך פירושם נשאר בלתי־מובן. לחילופין, ייתכן ולראשי התיבות יש משמעויות נוספות ולא רק זו שאליה התכוון מחבר הטקסט, וכל משמעות עלולה לשנות את פירוש הטקסט. לכן חשוב ביותר לזהות את המשמעות הנכונה של ראשי התיבות, בהקשר הנתון, על מנת להבין את הטקסט.

למיטב דעתנו, כל השיטות הקיימות לבנייה אוטומטית מקורפוסים של מילונים של ראשי תיבות מטפלות אך ורק בראשי תיבות מקומיים, שמשמעותם נמצאת באותו המסמך, בדרך־כלל בסמוך לשימוש הראשון בראשי התיבות, ופעמים רבות בסוגריים. לדוגמה, **א"א** הם ראשי תיבות מקומיים בכל אחד מהטקסטים הבאים:

- "אבא אבן (א"א) היה דיפלומט ופוליטיקאי."
- "התפילה מימי־הביניים מתייחסת לא"א (אברהם אבינו)."
- "ראשי התיבות של אנרגיה אטומית הם א"א."
- "היא טופלה באלכוהוליסטים אנונימיים. עכשיו היא מדריכה קבוצות של א"א."

בניגוד לזאת, ראשי־תיבות גלובלייס אינם מלווים במשמעויותיהם באותו המסמך, והם מובאים תחת ההנחה (השגויה לעתים) שהקורא מכיר את משמעות ראשי התיבות. ראשי התיבות הגלובליים האלה מהווים אתגר גדול יותר לבעיית הזיהוי.

מחקרנו מתבסס על אוסף גדול של טקסטים חופשיים (קורפוסים) בעברית המורכבים מעיתונות ישראלית, תמלילי דיונים מהכנסת, פרקים מספרות יפה, והגרסה העברית של האנציקלופדיה האינטרנטית ויקיפדיה. סך הכול, האוסף שלנו כולל 215 אלף מסמכים המכילים 77 מיליון תמניות (tokens) שמתוכן 950 אלף תמניות שונות (types), לא כולל מספרים, פיסוק ומילים לועזיות.

התיזה מקדמת את מצב המחקר בכמה אופנים:

- **שיטה לבניית מילונים של ראשי תיבות ומשמעויותיהם עם מידע הקשרי, הכוללים ראשי תיבות גלובליים:** פיתחנו שיטה חדשה לחילוץ אוטומטי של ראשי תיבות והמשמעויות שלהם מתוך קורפוסים של טקסט חופשי, על מנת לבנות מילון של ראשי תיבות ושל משמעויות שמחוזק באמצעות ההקשר. השיטה כוללת במפורש ראשי תיבות גלובליים, ועל כן זהו המחקר הראשון, למיטב ידיעתינו, שמטפל בקבוצה החשובה הזו של ראשי תיבות. מילונים שנבנים באמצעות השיטה הזו הם קלים לעדכון, וניתן ליצור אותם ליישם אותם עבור תחומים פרטניים.

  בשלב הראשון, השיטה זיהתה את כל ראשי תיבות ו־$n$־גרמים (רצף של $n$ מילים עוקבות) בקורפוסים. בשלב השני, התאמנו בין ראשי התיבות ל־$n$־גרמים כך ש־$n$־גרם וראשי תיבות יותאמו אם יש ביניהם אותיות משותפות בדפוסים שלפיהם נוצרים ראשי תיבות. חישבנו מספר מאפיינים בלשניים שקשורים לראשי תיבות, ה־$n$־גרם והזוג. אימנו מסווג מבוסס על שיטות של למידת מכונה כדי להבדיל

בין זוגות נכונים ולא נכונים. המסווג שלנו הגיע לדיוק של 0.82 ואחזור של 0.81 על קורפוסים בעברית.

- **משאב חדש בשפה העברית:** יישמנו את השיטה שלנו לבניית מילונים לקורפוסים עבריים, ויצרנו מילון ראשי תיבות עברי חדש, המתאים ליישומי עיבוד שפות טבעיות. אמנם קיימים כבר מילונים כאלו, אבל המילון שלנו גדול ומקיף יותר, ומכיל גם מידע הקשרי חשוב להתרת רב־משמעות של פירושי ראשי תיבות בטקסטים.
- **התרת רב משמעות של ראשי תיבות בעברית:** על מנת לבחון באופן נייטרלי את המילון שלנו, יישמנו אותו לבעיית התרת רב־משמעות של ראשי תיבות בהקשר נתון, והשגנו ביצועים מוצלחים יותר ביחס למילונים שנבנו בשיטות קיימות.

  במילון שלנו נתנו לכל ראשי תיבות מספר אפשרויות לפענוח שדורגו על פי מידת התאמתן בקונטקסט. ב־72.46% מהמקרים, המשמעות הראשונה שבמילון שלנו הייתה המשמעות הנכונה בקונטקסט, וכך הפחתנו את אחוז הטעיות ב־17.86% לעומת הסטנדרט החזק (מילון ראשי תיבות קיים שנערך ידנית ושאליו הוספנו את מספרי גימטרייה שכתובים כראשי תיבות, כגון, **י"ד** = **14**). בנוסף, ב־85.03% מהמקרים הזיהוי הנכון של ראשי התיבות היה שייך למילון שלנו–אם כי היה מדורג אחרי משמעויות לא נכונים לאותו הקשר. בכך הפחתנו את הטעויות ב־13.79% לעומת הסטנדרט. גם השווינו את הביצועים של המילון שלנו לסטנדרט אחר (מילון הבנוי מראשי תיבות מקומיים בסוגריים) ושם ההפחתה הייתה ב־52.38%.
- **תובנות בלשניות הנוגעות לראשי תיבות בעברית:** חקרנו לראשונה מזווית סטטיסטית תכונות בלשניות של ראשי תיבות בעברית ושימושם. תובנות אלו תועלנה לבלשנים, לחובבי השפה העברית, ולמפתחים של מערכות עבריות לעיבוד שפות טבעיות, שרוצים שעבודתם תיושם טוב יותר עבור ראשי תיבות.

  חלק מהתובנות הבלשניות היו:

  - ראשי תיבות הם הרבה יותר נפוצים בעיתונות ובדיוני הכנסת מאשר בספרות יפה ובאנציקלופדיות.
  - קבוצת ראשי התיבות אינה סגורה היא גדלה מיום ליום וככל שמוסיפים עוד קורפוסים מוצאים ראשי תיבות חדשים. תובנה זו מחזקת את הצורך בדרכים אוטומטיות וקלות לעדכון לבניית מילון ראשי תיבות.
  - 99% מראשי התיבות הם בני 2-6 אותיות ומשמעויותיהם הן בנות 2-5 מילים.
  - כמחצית מראשי תיבות נוצרים מהאותיות הראשונות של כל מילה במשמעות, לדוגמה **אעפ"כ** = **אַף עַל פִּי כֵן**. למרות זאת, יש והמשמעות נוצרת מיותר מאות אחת ממילה, כגון, **כדוה"א** = **כְּדִור הָאָרֶץ**. במספר קטן של ראשי

תיבות דילגו על אותיות שימוש שבהתחלת המילה או אפילו דילגו על מילים שלמות, לדוגמה **בג"צ** = **בֵּית דין גָּבוה לצֶדֶק**.

- אין יחס משמעותי בין השכיחות של ראשי התיבות ושכיחות משמעויותיהם. לדוגמה, **ש"ח** הוא שכיח מאד אבל **שֶקֶל חָדש** הוא צירוף נדיר; **את"א** הוא צירוף נדיר, ו**אוניברסיטת תֵל אָביב** הוא יותר שכיח. לבסוף, ל**אחה"צ** ול**אַחֵר הצָהריים** יש שכיחות דומה.
- הסיומות הנפוצות ביותר הן +יס (צורן לרבים זכר), +ית (צורן לנקבה), ו+י (צורן ההופך שם עצם לשם תואר, לדוגמה **צה"לי** = **קשור לצה"ל**; או קניין גוף ראשון יחיד, לדוגמה **צה"לי** = **צה"ל שלי**).

את המחקר שלנו ניתן להמשיך במספר כיוונים. באופן מיידי ניתן ליישם את הטכניקות שלנו לסוגות ספציפיות, לדוגמה טקסטים צבאיים, טקסטים של תאגידים וטקסטים הלכתיים. ניתן גם ליישם את השיטות שלנו לשפות נוספות–בפרט לשפות הדומות לעברית ושיש להן מורפולוגיה מורכבת ו/או שעדיין אין להן כלים רבים לעיבוד שפות טבעיות. גם למדנו שהרבה מהמשמעויות של ראשי תיבות הם ביטויים הבנויים ממספר מילים (multiword expressions, כגון, **אעפ"כ** = **אַף עַל פִּי כֵן**, או ישויות (named entities, כגון, **ק"ש** = **קְרית שְמונה**). אנו חושבים שהמחקר שלנו יכול לעזור בזיהוים. לבסוף, אם כי התרת רב משמעות היא דרך הטבעית ביותר להעריך את המילון שבנינו, ניתן היה לבחון את המחקר על ידי מדידת שיפור הביצועים של שליפת מידע או תרגום מכונה.