Proceedings of the Informing Science + Information Technology Education Conference

An Official Publication of the Informing Science Institute
InformingScience.org

InformingScience.org/Publications

June 30 – July 4, 2019, Jerusalem, Israel

EXTRACTING AND TAGGING UNSTRUCTURED CITATION OF A HEBREW RELIGIOUS DOCUMENT

Dror Mughaz*	Dept. of Computer Science, Bar-Ilan University, Ramat-Gan, Israel and Dept. of Computer Science, Lev Academic Center, Jerusalem, Israel	myghaz@gmail.com
Yaakov HaCohen-Kerner	Dept. of Computer Science, Lev Academic Center, Jerusalem, Israel	kerner@jct.ac.il
Dov Gabbay	Dept. of Computer Science, Bar-Ilan University, Ramat-Gan, Israel	dov.gabbay@kcl.ac.uk

* Corresponding author

ABSTRACT

Aim/Purpose	Finding and tagging citation on an ancient Hebrew religious document. These documents have no structured citations and have no bibliography.
Background	We look for common patterns within Hebrew religious texts.
Methodology	We developed a method that goes over the texts and extracts sentences containing the names of three famous authors. Within these sentences we find common ways of addressing those three authors and with these patterns we find references to various other authors.
Contribution	This type of text is rich in citations and references to authors, but because there is no structure of references it is very difficult for a computer to automatically identify the references. We hope that with the method we have developed it will be easier for a computer to identify references and even turn them into hyperlinks.
Findings	We have provided an algorithm to solve the problem of non-structured citations in an old Hebrew plain text. The algorithm definitely was able to find many citations but it has missed out some types of citations.

Accepted by Executive Review by Editor Eli Cohen | Received: February 18, 2019 | Revised: February 27, 2019 | Accepted: February 29, 2019.
Cite as: Mughaz, D., HaCohen-Kerner., Y., & Gabbay, D. M. (2019). Extracting and tagging unstructured citation of Hebrew religious document. *Proceedings of the Informing Science and Information Technology Education Conference*, *Jerusalem, Israel,* pp. 461-473. Santa Rosa, CA: Informing Science Institute. https://doi.org/10.28945/4345

Impact on Society	When the computer recognizes references, it will be able to build (at least partially) a bibliography that currently does not exist in such texts at all. Over time, OCR scans more and more ancient texts. This method can make people's access and understanding much.
Future Research	After we identify the references, we plan to automatically create a bibliography for these texts and even transform those references into hyperlinks.
Keywords	citations, text-mining, information extraction, Hebrew

INTRODUCTION

Extraction of citation and using them became very widespread. Citations have great potential to provide important information to researchers in various domains, such as academic, legal and religious. Currently, computerized corpora and search engines enable accurate extraction of citations; as a result, citation analysis is of prime importance.

The goal of this study is finding and tagging citations on an ancient Rabbinic Responsa (rabbinic scholars, also called poskim in Plural and posk in singular, that wrote answers in response to Jewish legal questions) documents. The majority of the studies on the subject of references conducted on the texts of scientific articles written in English. In scientific articles, there are defined and permanent references, typically with parentheses: citation with numbers (e.g., [1]), mixed symbols such as [Cohen 1998] (Harvard-style citations) or [Cohen98] or in footnotes, and at the end of each article, there is a bibliography. On the other hand, the Semitic languages and especially Rabbinic Responsa use citations differently than in academic documents.

Semitic languages are fairly dissimilar from Latin languages; thus, Hebrew and especially responsa documents will have different processing from the English language because: (1) texts written in Latin languages are written from left-to-right, while those written in Hebrew are written from right-to-left (Wintner, 2004); (2) there is no reference section in rabbinical texts as in academic papers; (3) Natural Language Processing (NLP) on combine texts of Hebrew, Aramaic and Yiddish, has barely been done to date (4) one sentence can contain words from three languages Hebrew Aramaic and Yiddish, which adds to the complex morphology of Semitic languages (e.g., words can exist with quite a few forms of prefixes "and when in …", "and when …", "and …", "when …", "in …"). One of the results of this complex morphology is ambiguous words (HaCohen-Kerner, Kass & Peretz, 2010); (5) the rabbinic references are not dated (HaCohen-Kerner, Schweitzer & Mughaz, 2011). (6) in Rabbinic texts, which are written in Hebrew Aramaic and Yiddish, contain a lot of acronyms and abbreviations than in regular Hebrew text (HaCohen-Kerner, Kass & Peretz, 2004). HaCohen-Kerner et al. (2004) show that there are 40,000 abbreviations in Hebrew, compared to 17,000 in English. Research done in 2013 (HaCohen-Kerner, Kass & Peretz, 2013) exhibits that the disambiguation manually of an acronym is a lot of time-consuming process, and it is very hard task even for a professional. Because of these characteristics of the responsa files make it very difficult to identify references.

RELATED WORKS

There have been various researches on citations in information retrieval (IR). However, none of them was in Hebrew or in rabbinic literature.

Garfield (1965) was the first to cope with the issue of automatic extraction, analysis and indexes production of citation, from scientific documents. He finds that citation can be used to discover emerging fields of science, analyze research trends, and to know the reputation of an article. Garfield also lists multiple reasons for citations.

Bergmark (2000) examine the issue of automatically reference linking online scholarly documents and introduce few algorithms for extracting metadata from online documents and linking full-text

documents together. She examined 66 papers and reported on average automatic extraction of online references of 83.1%.

Berkowitz and Elkhadiri (2004) developed a system that download, classify and index documents from the internet. Their system is also able to pull out titles from these documents and the names of the authors. They reported recall of 25.96% for accurate extraction of author names. Giuffrida, Shek and Yang (2000) used a knowledge-based system to extract meta-information from a set of 100 PostScript files of a computer science articles. They reported accuracy of 87% for extraction of author names. Seymore, McCallum and Rosenfeld (1999) used hidden Markov models for a similar task, reporting accuracy of 93.2% for author name extraction from a limited corpus of computer science articles.

In the last fifteen years developments (e.g., search engines and computerized datasets) allow big scale and very good citations extraction. Thus, the interest in citation analysis increased. The availability of text corpora and the new abilities of search engines provides good source of data for analyzing citation. Tan, Kan, and Lee (2006) present an approach to author disambiguation for the results of web searches. Teufel, Siddharthan, and Tidhar (2006) extracted citations and their context and used it for citations classification to their citation function (the reason for citing a given paper).

Some study has been done about retrieval performance achieved using paper terms. Bradshaw (2003) uses tokens from a fixed window around citations. Dunlop and van Rijsbergen (1993) use the abstracts of citing papers. Ritchie, Teufel, and Robertson (2008) show that document indexing based on combinations of expressions used by citing documents and expressions from the article itself give better retrieval performance than indexing only by document terms. Ritchie, Robertson, and Teufel (2008) examine how to select text from around the citations for extract good index terms for improve retrieval effectiveness.

Liebeskind (2009) used the taxonomy of text categorization via using the dice coefficient. Liebeskind extended every category using the top-k co-occurring word that got the top dice score. For all of the enlarged categories, they built a vector and computed the cosine similarity between it and the documents` vector. In a later work, Liebeskind, Dagan, and Schler (2012, 2016, 2018) built the thesaurus to an olden Hebrew-Aramaic language to help people which use Modern Hebrew to find information in this olden language. They developed an algorithm for building a dictionary for the thesaurus in morphologically-rich languages.

Mughaz, Fuchs, and Bouhnik (2018) classify short Hebrew texts according to their opinion. The corpus they used consists of short product reviews which were parsed into individual sentences. They applied the SVM algorithm on a combination of both unigrams and bigrams. Then they applied feature selection according to weights of the features by removing the features ranked less than 0.1 . They tested the pruned features on SVM with a linear kernel and Bayesian Logistic Regression which yield 92.6% and 92.4%, respectively.

Another works that are related to document classification and address the challenges of Hebrew involve the classification of Hebrew-Aramaic documents according to style (Koppel, Mughaz, & Akiva, 2006; Mughaz, 2003); authorship verification, including forgers and pseudonyms (Koppel, Mughaz, & Akiva, 2003; Koppel, Schler, & Mughaz, 2004); and classification of texts according to their ethnic origin and their historical period (HaCohen-Kerner, Beck, Yehudai & Mughaz, 2006, 2008; HaCohen-Kerner, Mughaz, Beck & Yehudai, 2010; HaCohen-Kerner, Beck, Yehudai, Rosenstein & Mughaz, 2010).

HaCohen-Kerner et al. (2011) used six machine learning techniques for identifying citations. To achieve this goal they used four feature types, n-gram, stop word-based, quantitative and orthographic, and tested them separately and together. The best results were by combination of the four feature sets. Their study recognized if a sentence included a citation, but, they did not recognize the citation itself.

Mughaz, HaCohen-Kerner, and Gabbay (2015) extracted time related key-phrases from rabbinical texts. They found that many of the sentences that hold time-related key-phrases are usually contain also rabbinic names. They applied and presented semi-automatic method that boosts the extraction of time-related key-phrases. In their works, Mughaz, HaCohen-Kerner, and Gabbay (2010, 2014, 2017, 2019) improved the previous method and used time-related phrases and references in order to date texts. The dating they suggested could help identify ancient anonymous texts and even help identify edited texts.

We are not aware of any other work on this subject in modern Hebrew let alone ancient Hebrew.

FINDING THE CITATION

As mentioned earlier, the majority of the studies on the subject of references conducted on the texts of scientific articles written in English. In scientific articles, there are defined and permanent references, and at the end of each article, there is a bibliography. On the other hand, in the texts of responsa there is no defined structure for references, and, more so, there is no bibliography. These two characteristics of the responsa files make it very difficult to identify references.

MARKING REFERENCES BY USING EXTRACTED TEMPLATES (BY SEARCHING FOR THREE IMPORTANT AUTHORS)

In order to deal with the problem of no defined reference structure, we decided to look for a set of templates.

The general algorithm

1) Search for sentences with names of important *poskim*.
2) Extract common patterns of citations.
3) Collect names/nicknames/acronyms/abbreviations/book-names of *poskim* (manually).
4) Replace the common patterns of citations with a unique string (for every *posek*) in order to evaluate the results.

Stage 1 of the algorithm:
In order to find patterns of references, we searched the Responsa Project at Bar-Illan University (n.d.) for the names of three important poskim: Rashi, the Rambam and Rabbi Yosef Karo. From the search results of the Responsa Project, we extracted sentences containing the names of these three halachic authorities.

Stage 2 of the algorithm:
From the sentences of stage 1, we extracted 22 different common patterns of references to a posek.

In table 1, what appears in Bold-italic is the name of the author, and what appears in the bold is the name of the book. In the column containing the template, we marked ***yyy*** as the author's name and **zzz** as the name of the book.

Table 1 shows examples of the patterns addressing Rashi (an 11th-century posek). The first 22 patterns are divided as follow, (1) 12 patterns with the name of the posek and the name of the book, (2) 9 templates with the name of the book only, and (3) one pattern with the name of the posek only (Table 1 lines 1-22). Table 1 lines 23-27 shows a different writing of Rashi's name, and line 28 shows another book that Rashi wrote.

Table 1: examples and templates with translation of references patterns

#	Template	Example in Hebrew	Translation of the Hebrew example
1	*yyy* ז"ל בספר ה-zzz	**רש"י** ז"ל בספר **הפרדס**	***Rashi*** of blessed memory in the **Pardes** book
2	שם *yyy* שכתב בספר ה-zzz	שם **רש"י** שכתב בספר **הפרדס**	In the name of ***Rashi*** who wrote in the **Pardes** book
3	בספר ה-zzz השלם ל-*yyy*	בספר **הפרדס** השלם ל**רש"י**	In the complete **Pardes** book of ***Rashi***
3	ב-*yyy* בספר ה-zzz	ב**רש"י** בספר **הפרדס**	In ***Rashi*** in the **Pardes** book
4	שם ה-*yyy* בה-zzz	שם ה**רש"י** ב**הפרדס**	In the name of ***Rashi*** in the **Pardes**
5	שו"ת ה-zzz ל-*yyy*	שו"ת **הפרדס** ל**רש"י**	The **Pardes** responsa of ***Rashi***
6	מ"ש *yyy* בה-zzz	מ"ש **רש"י** ב**הפרדס**	Which wrote ***Rashi*** in the **Pardes**
7	*yyy* בספר ה-zzz	**רש"י** בספר **הפרדס**	***Rashi*** in the **Pardes** book
8	ספר ה-zzz ל-*yyy*	ספר **הפרדס** ל**רש"י**	The **Pardes** book of ***Rashi***
9	ס' ה-zzz ל-*yyy*	ס' **הפרדס** ל**רש"י**	The **Pardes** b' of ***Rashi***
10	*yyy* בעל ה-zzz	**רש"י** בעל **הפרדס**	***Rashi*** the possessor of the **Pardes**
11	*yyy* בה-zzz	**רש"י** ב**הפרדס**	***Rashi*** in the **Pardes**
12	ספר zzz הגדול ל-*yyy*	ספר **פרדס** הגדול ל**רש"י**	The great **Pardes** book of ***Rashi***
13	שם ספר ה-zzz	שם ספר **הפרדס**	In the name of the **Pardes** book
14	שם ס' ה-zzz	שם ס' **הפרדס**	In the name of the **Pardes** b'
15	ספר zzz הגדול	ספר **פרדס** הגדול	The great **Pardes** book
16	zzz הגדול	**פרדס** הגדול	The great **Pardes**
17	ספר ה-zzz	ספר **הפרדס**	The **Pardes** book
18	שו"ת ה-zzz	שו"ת **הפרדס**	The **Pardes** responsa
19	בעל ה-zzz	בעל **הפרדס**	The possessor of the **Pardes**
20	בשם ה-zzz	בשם **הפרדס**	In the name of the **Pardes**
21	ס' ה-zzz	ס' **הפרדס**	The **Pardes** b'
22	בשם *yyy*	בשם **רש"י**	In the name of ***Rashi***
23	*yyy* (the usual name)	**רש"י**	***Rashi***
24	*yyy* (name1)	***שלמה יצחקי***	***Shlomo Yitzhaki***
25	*yyy* (name2)	***שלמה בן רבי יצחקי***	***Shlomo*** son of Rabbi ***Yitzhaki***
26	*yyy* (name3)	***שלמה ב"ר יצחקי***	***Shlomo*** (acronym) son of Rabbi ***Yitzhaki***
27	*yyy* (name3)	***שלמה בן יצחקי***	***Shlomo*** son of ***Yitzhaki***
28	zzz (another book)	ספר **האורה**	The **Orah** book

A very common way of relating to rabbinic texts is to refer to the name of the book without reference to the author at all, i.e., the name of the book is the author's "name." A very prominent example is the reference to the "Shulchan Aruch." The "Shulchan Aruch" is the main halakhic book that was accepted by almost all Jews nowadays. The author of the "Shulchan Aruch" is Rabbi Yosef Karo, and many times when an author wishes to relate to Rabbi Yosef Karo, he merely writes "the

Shulchan Aruch" or "the master of the Shulchan Aruch" and everyone knows that Rabbi Yosef Karo is being referred to.

Stage 3 of the algorithm:
We manually collected 339 different authors with different forms of writing of their names (e.g., "Rabbi Yosef Karo" and "R' Yosef Karo"). For each author, we collected the books he composed, with the various forms of writing of the books' names (e.g., one book by Rabbi Yosef Karo with two different forms of writing, "Shulchan Aruch" = "שולחן ערוך" and "Shu"a" = "שו"ע"; an example of another book by Rabbi Yosef Karo is "Kesef Mishnah").

Stage 4 of the algorithm:
Above, we noted that a posek could write several different books, and could be referred to by different names or variations of his name. Each posek's citation pattern can be expanded to many other citation patterns by replacing the name of the author or his book by each one of his other names (e.g., different spellings, full names, short names, surnames, family names and nicknames with/without title), acronyms, and abbreviations. In order to do this, we took all the templates that appear in table 1 in rows 1 - 22; each yyy was replaced by a different form of the author name, and each zzz was replaced by a different form of the name of the book. For example, the book Shulchan Aruch can be written in two different forms (1) Shulchan Aruch, and (2) the acronym Shu"a. When mentioning the name of the author of the Shulchan Aruch, Rabbi Yosef Karo, his name can be written in two different ways: (1) Rabbi Yosef Karo, (2) R' Yosef Karo. In this example, we see that there are two different types of writing of the book's name and two different forms of author/Posek writing. In this example, we conclude that one pattern generates four different strings. See table 2.

Note that since there is no defined structure for references to rabbis, there are certainly references that we missed.

Table 2: Example of two different name writing of the same author and two different book-name writing of the same book for one template

Template	String in Hebrew that associate with the template	Translation of the Hebrew strings
-ספר הzzz -לyyy	ספר **השולחן ערוך** לרבי **יוסף קארו**	The **shulhan a'aruch** book of *Rabbi Yosef Karo*
	ספר **השו"ע** לרבי יוסף קארו	The **shu"a'a (acronym)** book of *Rabbi Yosef Karo*
	ספר **השולחן ערוך** לר' יוסף קארו	The **shulhan a'aruch** book of *R' Yosef Karo*
	ספר **השו"ע** לר' יוסף קארו	The **shu"a'a (acronym)** book of *R' Yosef Karo*

Processing the templates in Table 1, with all 339 authors, including the different forms of the author's names, and the names of the books, including the different forms of the book names, gave us 25,970 different strings representing references to different authors and books.

In order to evaluate the results easily and also to know whether we have covered the entire reference we did the following: For each author, and all the books he wrote, we created a unique string that does not appear at all in our data set. In order to create a unique string for each posek, we used strings with a linguistic error (which does not exist in the data set) that can be found by regular expression easily. See the partial representative example in Table 3; each of the unique strings begins with the letter `ף` (which linguistically appears only at the end of a word) and also ends with the letter `ף` (even when there is no linguistic need).

Table 3: Partial example of two (of 339) different authors with the unique string

#	Translation of the Hebrew example	Example in Hebrew	Unique string
1	In the name of ***Rabbi Ovadya Yosef*** who wrote in the **Hazon Ovadya** book	שם ***הרב עובדיה יוסף*** שכתב בספר **חזון עובדיה**	ויעובדיוסף
2	In the complete **Hazon Ovadya** book of ***Rabbi Ovadya Yosef***	בספר **חזון עובדיה** השלם ***להרב עובדיה יוסף***	ויעובדיוסף
3	...	...	...
4	***Rabbi Ovadya Yosef*** the possessor of the **Hazon Ovadya**	***הרב עובדיה יוסף*** בעל **חזון עובדיה**	ויעובדיוסף
5	***Rabbi Ovadya Yosef*** in the **Hazon Ovadya**	***הרב עובדיה יוסף*** ב**חזון עובדיה**	ויעובדיוסף
6	...	...	...
7	The **Hazon Ovadya** b'	ס' **חזון עובדיה**	ויעובדיוסף
8	...	...	...
9	In the name of ***Rabbi Ovadya Yosef*** who wrote in the **Taharat Habait** book	שם ***הרב עובדיה יוסף*** ב**טהרת הבית**	ויעובדיוסף
10	Which wrote ***Rabbi Ovadya Yosef*** in the **Taharat Habait**	מ"ש ***הרב עובדיה יוסף*** ב**טהרת הבית**	ויעובדיוסף
11	...	...	...
12	In the complete **Taharat Habait** book of the genius ***Rabbi Ovadya Yosef***	בספר **טהרת הבית** השלם לגר***ע"י***	ויעובדיוסף
13	...	...	...
14	The **Taharat Habait** book	ספר **טהרת הבית**	ויעובדיוסף
15	...	...	...
16	In the name of ***Rambam*** who wrote in the **Yad Hhazakah** book	שם ***רמב"ם*** שכתב בספר **היד החזקה**	וירמבמיף
17	In the complete **Yad Hhazakah** book of the ***Rambam***	בספר **היד החזקה** השלם ל***רמב"ם***	וירמבמיף
18	Which wrote ***Rambam*** in the **Yad Hhazakah**	מ"ש ה***רמב"ם*** ביד החזקה	וירמבמיף
19	...	...	...
20	The possessor of the **Yad Hhazakah**	בעל **היד החזקה**	וירמבמיף
21	In the name of the **Yad Hhazakah**	שם **היד החזקה**	וירמבמיף
22	...	...	...
23	***Rambam*** in the **Mishne Torah** book	***רמב"ם*** בספר **משנה תורה**	וירמבמיף
24	***Rambam*** the possessor of the **Mishne Torah**	***רמב"ם*** בעל ה**משנה תורה**	וירמבמיף
25	...	...	...
26	the possessor of the **Mishne Torah**	בעל ה**משנה תורה**	וירמבמיף
27	**Mishne Torah** book	ספר **משנה תורה**	וירמבמיף

Table 3 contains examples of references to two authors, the Rambam (a 12th century posek) and Rabbi Ovadia Yosef (a 20th century posek). The Rambam wrote a book called "the Yad Hahazakah," also known as "the Mishne Torah" (different references to the same book). Rabbi Ovadia Yosef wrote a book called "Hazon Ovadia" and another book called "Taharat HaBayit." All the references to the Rambam`s various books or his name were combined into a unique string representing the Rambam, and all the references to Rabbi Ovadia Yosef's various books or his name were combined into a unique string representing him as explained earlier. We ran a program on all the files of all our authors that replaced all the references to a particular book or author with a unique string representing that same author.

DATA SET AND RESULTS

The texts of the studied corpus were downloaded from Bar-Ilan University's Responsa Project. The studied corpus holds 15,495 responsa written by 24 poskim, each responsa author has 643 files on average (see Appendix). The overall number of characters in the entire corpus is 127,683,860 chars, and the average count of chars for per file is 8,240 chars. These writers lived over a period of 229 years (1786–2015). These files contain citations; each citations pattern can be written in various other specific citations forms (HaCohen-Kerner et al., 2011).

Citations identification was done by comparing every word to a list of 339 known poskim and many of their books, as explained before. This list of 25,801 specific citations refers to the abbreviations, nicknames and names of these authors and their books. Basic references were gathered (table 1) and all other references were composed according to them.

Table 4. Details about the Citation in the data set

total citations	Average Citation per author
404116	16838.2

Table 4 contains the number of references that our system has identified. The identified citations that were marked/substitute are the names/abbreviations/acronyms /book-names of a posekim and the number of it them very large.

PRACTICAL IMPLICATIONS

Search engines can refer to references/citations appearing in a plain text as hyperlink. The problem with such references/citations in a plain text is that they are not structured in such a way that it is easy for a computer to convert them to hyperlinks. The meaning of search engines' addressing of certain parts of plain text as hyperlink is significant. The search engines can use these hyperlinks to build a graph that links between different web-sites or document and use it as additional parameter in their results. If the search engines will construct the graph to be directed graph, the graph can also get a partial meaning of the chronological order of which site/document appeared earlier. In the same line of thought it is possible to assess the time frame in which unknown texts were written or to estimate when certain authors lived. In addition, search engines will be able to use citations to know which site or document is important and influential by the amount of citations it got, in terms of fan-in graphs.

These unique plain text parts, i.e., references and citations, can give a different perspective on web content. Thus, extraction of unstructured citation is important and challenging task and it is the first step toward convert citations to hyperlinks.

LIMITATIONS AND ERROR ANALYSIS

One problem with this method is the manual search element. To find the patterns appearing in Table 1 in all the texts, we need all the names of all authors. However, each author has many different

"names", that is, his full name, the acronym of his name, referring to him through the name or the names of the books he wrote, etc. For example see Table 2. Therefore, we must manually search for the different ways the same author is referred to. An automatic system cannot know, for example, that the Rambam is sometimes referred to as "the Great Eagle" or that the "Ben Ish Chai" is Rabbi Yosef Chaim of Baghdad.

The second problem is that there are authors that do not appear on our 339 list of writers. Because we rely on a list of authors, a non-listed author will not be recognized by the system as a citation.

The third problem is that there are cases where we failed to discover the entire reference. These situations happen with citations that contain indexes for sections within a certain part of a particular book, examples below:

The bold are marked references and the italicized are parts of the unmarked references. (In these examples, we can see the length and the complexity of a citation in this kind of text.)

(1) שו"ת ב"ח החדשות סימן מז

... לא משמע הכי בדברי **הרמב"ם** *פ"א מהל' ת"ת* דאע"פ שכת' כשם שחייב אדם ללמוד את בנו ...

(1) Responsa the New B"h (Bait Hadash) chapter 47

… this is not the meaning the words of the **Ramba"m** (acronym: Rabbi Moshe ben Maimon) *C"A of Hil` T"T* (acronym and abbreviation: *chapter A of the halachot* (religious laws) *Talmud Torah*) even though he wrote as a man must teach his son …

(2) שו"ת חתם סופר חלק א (אורח חיים) סימן צ

... ומכ"ש לפמ"ש **הרמב"ם** *פ"א מה' שבת הלכה י"ב* עפ"י דרך של ה"ה שם ...

(2) Responsa Chatam Sofer Part 1 (Orach Chaim) chapter 90

… all the more so according to what the **Ramba"m** (acronym: Rabbi Moshe ben Maimon) *c11 of Hil` shabat halacha 12* (acronym and abbreviation: *chapter 11 of Hilchot shabat halacha 12*) according to the Maggid Rabbi there …

(3) שו"ת בית יצחק אבן העזר ב סימן צב

... א"כ לפמ"ש **הטו"ז** *בח"מ סי' ל"ד* דשכר טירחות ההליכ' רשאין ליקח ...

(3) Responsa Beit Itzhak Even HaEzer Part 2 chapter 92

… if so according to what **Tu"Z** in *H"M cha` 34* (acronym and abbreviation: Turi Zahav in *Choshen Mishpat chapter 34*) wrote that they are entitled to take the walking fee …

CONCLUSIONS

In this study, we presented the problem of non-structured citations in a plain text. We explained the important potential of such citations to search engines, to date texts and writers, and to identify of edited texts. We explained how search engines can improve using these citations. In the context of the Hebrew language we have shown that this problem is much greater in Hebrew texts in general, and especially in rabbinical texts. We have provided an algorithm to solve the problem. The algorithm has indeed been able to find many references but there is still much to be done in the field. The current state is far from an accepted solution. We need to strive for more complete and accurate solution of the non-structured citations tagging.

Because of the structure of the rabbinic texts and the way of the references in them, we can use expressions of respect and affection as another clue to refine the citations tagging and reduce the errors of marking them. More than that, we plan to overcome the problem of indexes in the citations of rabbinical texts by using deep learning techniques.

REFERENCES

Bergmark, D. (2000). *Automatic extraction of reference linking information from online documents*. Cornell University.

Berkowitz, E., & Elkhadiri, M. R. (2004). *Creation of a style independent intelligent autonomous citation indexer to support academic research*. Retrieved from http://cogprints.org/3661/1/ebmaics2004a.pdf

Bradshaw, S. (2003, August). Reference directed indexing: Redeeming relevance for subject search in citation indexes. In *International Conference on Theory and Practice of Digital Libraries* (pp. 499-510). Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-540-45175-4_45

Dunlop, M. D., & van Rijsbergen, C. J. (1993). Hypermedia and free text retrieval. *Information Processing & Management, 29*(3), 287-298. https://doi.org/10.1016/0306-4573(93)90056-j

Garfield, E. (1965, December). Can citation indexing be automated. In *Statistical association methods for mechanized documentation, symposium proceedings* (Vol. 269, pp. 189-192). Washington, DC: National Bureau of Standards, Miscellaneous Publication 269.

Giuffrida, G., Shek, E. C., & Yang, J. (2000, June). Knowledge-based metadata extraction from PostScript files. In *Proceedings of the fifth ACM Conference on Digital Libraries* (pp. 77-84). ACM. https://doi.org/10.1145/336597.336639

HaCohen-Kerner, Y., Beck, H., Yehudai, E., & Mughaz, D. (2006, October). Identifying historical period and ethnic origin of documents using stylistic feature sets. In *International Conference on Discovery Science* (pp. 102-113). Springer, Berlin, Heidelberg. https://doi.org/10.1007/11893318_13

HaCohen-Kerner, Y., Beck, H., Yehudai, E., & Mughaz, D. (2010). Stylistic feature sets as classifiers of documents according to their historical period and ethnic origin. *Applied Artificial Intelligence, 24*(9), 847-862. https://doi.org/10.1080/08839514.2010.514197

HaCohen-Kerner, Y., Beck, H., Yehudai, E., Rosenstein, M., & Mughaz, D. (2010). Cuisine: Classification using stylistic feature sets and/or name-based feature sets. *Journal of the American Society for Information Science and Technology, 61*(8), 1644-1657. https://doi.org/10.1002/asi.21350

HaCohen-Kerner, Y., Kass, A., & Peretz, A. (2004). Baseline methods for automatic disambiguation of abbreviations in Jewish law documents. In *Advances in Natural Language Processing* (pp. 58-69). Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-540-30228-5_6

HaCohen-Kerner, Y., Kass, A., & Peretz, A. (2010). HAADS: A Hebrew Aramaic abbreviation disambiguation system. *Journal of the American Society for Information Science and Technology, 61*(9), 1923-1932. https://doi.org/10.1002/asi.21367

HaCohen-Kerner, Y., Kass, A., & Peretz, A. (2013). Initialism disambiguation: Man versus machine. *Journal of the American Society for Information Science and Technology, 64*(10), 2133-2148. https://doi.org/10.1002/asi.22909

HaCohen-Kerner, Y., & Mughaz, D. (2010, August). Estimating the birth and death years of authors of undated documents using undated citations. In *International Conference on Natural Language Processing* (pp. 138-149). Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-14770-8_17

HaCohen-Kerner, Y., Mughaz, D., Beck, H., & Yehudai, E. (2008). Words as classifiers of documents according to their historical period and the ethnic origin of their authors. *Cybernetics and Systems: An International Journal, 39*(3), 213-228. https://doi.org/10.1080/01969720801944299

HaCohen-Kerner, Y., Schweitzer, N., & Mughaz, D. (2011). Automatically identifying citations in Hebrew-Aramaic documents. *Cybernetics and Systems: An International Journal, 42*(3), 180-197. https://doi.org/10.1080/01969722.2011.567893

Koppel, M., Mughaz, D., & Akiva, N. (2003). CHAT: *A system for stylistic classification of Hebrew-Aramaic texts*, The 3th Workshop on Operational Text Classification Systems.

Koppel, M., Mughaz, D., & Akiva, N. (2006). New methods for attribution of rabbinic literature., *Hebrew Linguistics, A Journal for Hebrew Descriptive, Computational, Applied Linguistics*, *57*, v-xviii.

Koppel, M., Schler, J., & Mughaz, D. (2004, January). Text categorization for authorship verification. In *Eighth International Symposium on Artificial Intelligence and Mathematics*. Fort Lauderdale, Florida, http://rutcor. rutgers. edu/~amai/aimath04/SpecialSessions/Koppel-aimath04.pdf.

Liebeskind, C., (2009). *Master work: Text categorization for large multi-class taxonomy*. Bar Ilan University. Department of Mathematics and Computer Science.

Liebeskind, C., Dagan, I., & Schler, J. (2012). *Statistical thesaurus construction for a morphologically rich language*. In * SEM 2012: The First Joint Conference on Lexical and Computational Semantics–Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation (SemEval 2012) (Vol. 1, pp. 59-64).

Liebeskind, C., Dagan, I., & Schler, J. (2016). Semiautomatic construction of cross-period thesaurus. *Journal on Computing and Cultural Heritage (JOCCH), 9*(4), 22. https://doi.org/10.1145/2994151

Liebeskind, C., Dagan, I., & Schler, J. (2018). Automatic thesaurus construction for modern Hebrew. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation* (LREC-2018).

Mughaz, D. (2003). *Classification of Hebrew Texts according to Style*. M.Sc. Thesis, Bar-Ilan University, Ramat-Gan, Israel, 2003.

Mughaz, D., Fuchs, T., & Bouhnik, D. (2018). Automatic opinion extraction from short Hebrew texts using machine learning techniques. *Computación y Sistemas, 22*(4). https://doi.org/10.13053/cys-22-4-3071

Mughaz, D., HaCohen-Kerner, Y., & Gabbay, D. (2014, November). When text authors lived using undated citations. In *Information Retrieval Facility Conference* (pp. 82-95). Springer, Cham. https://doi.org/10.1007/978-3-319-12979-2_8

Mughaz, D., HaCohen-Kerner, Y., & Gabbay, D. (2015, September). Key-phrases as means to estimate birth and death years of Jewish text authors. In *Semantic keyword-based search on structured data sources* (pp. 108-126). Springer, Cham. https://doi.org/10.1007/978-3-319-27932-9_10

Mughaz, D., HaCohen-Kerner, Y., & Gabbay, D. (2017). Mining and using key-words and key-phrases to identify the era of an anonymous text. In *Transactions on computational collective intelligence XXVI* (pp. 119-143). Springer, Cham. https://doi.org/10.1007/978-3-319-59268-8_6

Mughaz, D., Hacohen-Kerner, Y., & Gabbay, D. (2019). Text mining for evaluating authors' birth and death years. *ACM Transactions on Knowledge Discovery from Data* (TKDD), *13*(1), 7. https://doi.org/10.1145/3281631

The Responsa Project at Bar-Ilan University. (n.d.). Retrieved from http://www.biu.ac.il/ICJI/Responsa

Ritchie, A., Robertson, S., & Teufel, S. (2008, October). Comparing citation contexts for information retrieval. In *Proceedings of the 17th ACM Conference on Information and Knowledge Management* (pp. 213-222). ACM. https://doi.org/10.1145/1458082.1458113

Ritchie, A., Teufel, S., & Robertson, S. (2008, March). Using terms from citations for IR: Some first results. In *European Conference on Information Retrieval* (pp. 211-221). Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-540-78646-7_21

Seymore, K., McCallum, A., & Rosenfeld, R. (1999, July). *Learning hidden Markov model structure for information extraction*. In AAAI-99 workshop on machine learning for information extraction (pp. 37-42).

Tan, Y. F., Kan, M. Y., & Lee, D. (2006, June). Search engine driven author disambiguation. In *Proceedings of the 6th ACM/IEEE-CS Joint Conference on Digital Libraries* (pp. 314-315). ACM. https://doi.org/10.1145/1141753.1141826

Teufel, S., Siddharthan, A., & Tidhar, D. (2006, July). Automatic classification of citation function. In *Proceedings of the 2006 conference on empirical methods in natural language processing* (pp. 103-110). Association for Computational Linguistics. https://doi.org/10.3115/1610075.1610091

Wintner, S. (2004). Hebrew computational linguistics: Past and future. *Artificial Intelligence Review, 21,* 113-138.

APPENDIX: DATA SET INFORMATION

Full details about the data set

Author's name	Birth year	Death year	# of files	# of words	# of chars
Vozner Shmuel	1914	2015	1807	1,490,463	7,768,059
Yosef Ovadya	1920	2014	1283	4,578,049	22,933,473
Waldenberg Eliezer	1917	2006	1639	3,197,662	16,589,888
Auerbach Shlomo Zalman	1910	1995	229	793,706	4,087,592
Weiss Yitzchak	1902	1989	1468	2,311,927	11,695,021
Stern Bezalel	1911	1989	663	1,080,452	5,390,661
Feinstein Moshe	1895	1986	1831	2,306,526	11,959,224
Hadaya Ovadia	1890	1969	210	713,341	3,683,787
Ades Yaakov	1898	1963	131	310,585	1,604,218
Havita Rahamim	1901	1959	736	898,543	4,655,681
Herzog Yitzchak	1889	1959	190	430,259	2,210,586
Ben-Zion Meir Hai Uziel	1880	1953	374	899,617	4,621,414
Boimel Yehoshua	1880	1948	129	237,093	1,227,007
Baer Weiss Yitzchak	1873	1942	497	243,789	1,257,633
Kook Abraham Yitzchak	1865	1935	681	750,145	3,892,610
Allouch Faraji	1854	1921	112	205,258	1,069,460
Schwadron Sholom Mordechai	1835	1911	1574	1,657,860	8,560,084
Somekh Abdallah	1813	1889	86	80,508	412,486
Spektor Yitzchak Elchanan	1817	1896	301	1,159,019	5,843,696
Trunk Israel Yehoshua	1820	1893	281	132,257	689,598
Abuhatzeira Yaakov	1790	1880	146	177,411	917,682
Edery Abraham	1801	1874	119	176,849	918,564
Assad Yehuda	1794	1866	882	880,361	4,565,230
Birdugo Yaakov	1786	1843	126	218,402	1,130,206

BIOGRAPHY

Dror Mughaz is a doctoral student in the Department of Computer Science at Bar-Ilan University under the supervision of Professor Yaakov HaCohen-Kerner and Professor Dov Gabbay. Dror is also a lecturer at the Computer Science Department in the Jerusalem College of Technology. Dror is a co-author of 18 papers. Dror's current main research is in Text Mining of temporal issues. Other research fields that he works on are text clustering, citation extraction and analysis, word embedding, n-gram embedding, opinion mining, features and key-phrase extraction, data enrichment and author verification. Much of his research was done on Hebrew texts with a focus on rabbinical texts. He was a member of COST research actions of the EU.

Professor Yaakov HaCohen-Kerner is the Head of the Research Authority of the Jerusalem College of Technology. He also teaches at the Computer Science Department. Yaakov is a co-author and author of 87 papers. Yaakov's current main research is in Text Classification. Other research domains are text clustering, image and speech classification, author profiling, word completion and prediction, key-phrase extraction, citation extraction and analysis, plagiarism detection, and composition of chess and checkers problems. A few of these research domains are implemented on Jewish religious writings as well as scientific papers, social data, etc. He is or was a member of six COST research actions of the EU.

Professor Dov M. Gabbay is an Israeli logician. He is Emeritus Professor at Bar Ilan University in the Department of Computer Science, he is Augustus De Morgan Professor Emeritus of Logic at Logic, Language and Computation, Department of Computer Science, King's College London and Visiting Professor at the University of Luxembourg.

Gabbay has authored over five hundred research papers and over thirty research monographs. He is editor of several international Journals, and many reference works and Handbooks of Logic, including the Handbook of Philosophical Logic, the Handbook of Logic in Computer Science (with Samson Abramsky and T. S. E. Maibaum), and the Handbook of Artificial Intelligence and Logic Programming.