Key-Phrases as Means to Estimate Birth and Death Years of Jewish Text Authors

Dror Mughaz[1,2(✉)], Yaakov HaCohen-Kerner[2], and Dov Gabbay[1,3]

[1] Department of Computer Science, Bar-Ilan University,
5290002 Ramat-Gan, Israel
myghaz@gmail.com, dov.gabbay@kcl.ac.uk
[2] Department of Computer Science, Lev Academic Center,
9116001 Jerusalem, Israel
kerner@jct.ac.il
[3] Department of Informatics, Kings College London,
Strand, London WC2R 2LS, UK

Abstract. In this study, we try to determine the time-frame in which the author of a given document lived. The discussed documents are rabbinic documents written in the Hebrew, Aramaic and Yiddish languages. The documents are usually undated and do not contain a bibliographic section, which leaves us with an interesting challenge to determine the desired time-frame. To do this, we define a set of key-phrases and formulate various types of rules: "Iron-clad", Heuristic and Greedy constraints, to define the time-frame. These rules are based on key-phrases and key-words in the documents of the authors. Identifying the time-frame of an author can help us determine the generation in which specific documents were written, can help in the examination of documents, i.e., to conclude if documents were edited, and can also help us identify an anonymous author. We tested these rules on two corpuses of documents, which were authored by 12 and 24 rabbinic authors, respectively, and the results are promising.

Keywords: Hebrew-Aramaic documents · Key-phrases · Key-words · Knowledge discovery · Time analysis · Undated documents · Undated references

1 Introduction

Determining the time frame of a book or a manuscript and identifying an author are important and challenging problems. Time-related key-words and key-phrases with references can be used to date and identify authors. Key-phrases and key-words have great potential to provide great information in many domains, such as academic, legal and commercial. Thus, the automatic extraction and analysis of key-phrases and key-words is growing rapidly and gaining momentum. Web search engines, machine learning, etc. are based on key-phrases and key-words. As a result, features are extracted and learned automatically; thus, the analysis of key-phrases and key-words has enormous importance. Key-phrases and key-words are essential features not only of the needs of scientific papers or for industry and commerce but also of rabbinic

J. Cardoso et al. (Eds.): KEYWORD 2015, LNCS 9398, pp. 108–126, 2015.
DOI: 10.1007/978-3-319-27932-9_10

responsa (answers written in response to Jewish legal questions authored by rabbinic scholars). Key-phrases and words included in rabbinic text are more complex to define and to extract than key-phrases and key-words in academic papers written in English because: (1) Natural Language Processing (NLP) in Hebrew, Aramaic and Yiddish has been studied relatively little; (2) there is a mixture of the complex morphology of Hebrew, Aramaic and Yiddish. For example, key-phrases and key-words can be presented with different types of prefixes (e.g., "and …", "when …", "and when …", "in …", "and when in …"); (3) key-phrases and words in Hebrew, Aramaic and Yiddish texts may be ambiguous. Responsa texts present various interesting problems: (1) the morphology in Hebrew is richer than in English. Hebrew has 70,000,000 valid forms, while English has only 1,000,000 [1]. Declensions in Hebrew can be up to 7000 for one stem, while in English, there are only a few declensions; (2) responsa documents have a high rate of abbreviations (nearly 20 %) [2].

This research estimates the date of undated documents of authors using (1) the year (s) mentioned in the text, (2) "late" ("of blessed memory") key-phrases, (3) "rabbi" key-phrases, (4) "friend" key-phrases that are mentioned in the texts and (5) undated references of other dated authors that refer to the considered author or are mentioned by him. The assessments are with different degrees of certainty: "iron-clad", heuristic and greedy. The rules are based on key-phrases with and without references.

This paper is organized as follows: Sect. 2 gives background concerning the extraction and analysis of key-phrases and citation. Section 3 presents the boosting extraction key-phrases algorithm. Section 4 presents various rules of some degrees of certainty: "iron-clad", heuristic and greedy rules, which are used to assess writers' birth and death years. Section 5 presents the model description. Section 6 familiarizes the dataset, experiments, results and analysis. Section 7 includes the summary, conclusions and future works.

2 Related Research

Following the explosion of electronic information, there has been a growing need for extracting key-phrases and words automatically. Many studies have been made in this area for different purposes and from different perspectives. Key-words in documents allow for quick search on multiple large databases [3]. Key-words can also help to improve the NLP performance, as well as Information Retrieval performance in issues such as text summarization [4], text categorization [5], topic change during conversational text [6] and opinion mining [7].

Although key-words are important in many computer programs, there is still much to be done in this area, and the state-of-the-art methods underperform compared to other NLP core tasks [8].

There are several difficulties in extracting key-phrases and key-words. One is the length of the documents. In scientific articles, although there can only be approximately 10 key-words or key-phrases and approximately another 30 candidates in the abstract section, the rest of the article may contain hundreds of candidate key-phrases or key-words [9]. Moreover, key-words can also appear at the end of an article. If key-phrase or key-word appears at the beginning and at the end of an article, it indicates the importance of that key-phrase or key-word [10].

When documents are structured, key-words extraction is easier. For example, in scientific papers, most of the key-words appear in the abstract, in the introduction and in the title [11]. In other cases, key-phrases can be automatically extracted from web page text and from its metadata [12] for the purpose of advertisement.

Automatic extraction and analysis of references from academic papers was first proposed by Garfield [13]. Berkowitz and Elkhadiri [14] extracted writers' names and titles from articles. A knowledge-based system was used by Giuffrida et al. [15] to derive metadata, including writers' names, from computer science journal articles. Hidden Markov Models were used by Seymore et al. [16] to extract writer names from a limited collection of computer science articles. The use of terms leads to progress in the extraction of information. Selecting text before and after references to extract good index terms to improve retrieval effectiveness was done by Ritchie et al. [17]. Bradshaw [18] used terms from a fixed window around references.

In contrast with scientific articles, the documents we are working on are from the Responsa Project[1]; they are without any structural base, usually contain a mixture of at least two languages, and contain noise. Previous research on the Responsa Project dealt with text classification [19]. They checked whether classification could be done over the long axis of ethnic groups of authors with stylistic feature sets. HaCohen-Kerner and Mughaz [20] and Mughaz et al. [21] investigated in which era rabbis lived using undated Responsa, but they did not address the problem of how to extract time-related key-words or key-phrases. This article is a continuation research of this issue, i.e., determining when writers lived using key-phrases.

3 Semi-automatic Boosting Extraction of Key-Phrases

We want to automate the extraction of time-related key-phrases. We found that most of the sentences that contain time-related concepts (i.e., time-related words and phrases) to rabbinic literature (e.g., "late", "friend") are usually nearby rabbinic names/nicknames/acronyms/abbreviations/book-names. We developed a semi-automatic algorithm that boosts concepts extraction for extracting time-related concepts. The main idea is to extract sentences that contain names of rabbis so that the words and phrases that are nearby the rabbinic names are treated as the key-phrases (among others) that we look for. Now, we present a general description of our extraction algorithm.

3.1 The Algorithm

Notations:

TW – vector of Time-related Words.
RN – vector of Rabbinic Names.
n – number of iterations of the algorithm.

[1] Contained in the Global Jewish Database (The Responsa Project at Bar-Ilan University). http://www.biu.ac.il/ICJI/Responsa.

```
TRC – set of Time-Related Concepts starting with, e.g., year, life, colleague, era.
TW ← TRC //initiate TW vector with TRC set
For i = 1 to n, do:
  • Search for sentences that contain the last concepts added to TW.
  • Extract new rabbinic names from those sentences.
  • Add the new rabbinic names to RN.
  • Search for sentences that contain the last rabbinic names added to RN.
  • Extract time-related concepts from the new sentences:
      – Delete stop words.
      – Add the new time-related concepts to TW.
      – Add the new time-related words and phrases to TRC (with their frequency)
        and for the "old" time-related words and phrases only add their frequency.
        Sort TRC by the frequency of time-related words and phrases in decreasing
        order (normally, concepts have a large number of appearances).
  • Select from the table the most frequent time-related concepts.
```

3.2 Algorithm Results

After using the algorithm, we extract time-related key-words, key-phrases and acronyms (a partial list is shown in Table 1). We divide them into three Hebrew and Aramaic key-words and key-phrases sets:

Rabbi – addressing another person as a rabbi/master, i.e., there is overlap between the lifetime of one author and the lifetime of another who is referred to by the first author as rabbi.

Friend – addressing another person as a friend, i.e., there is a large overlap between the lifetime of one author and the lifetime of another who is referred to by the first author as a friend.

Late – addressing a person who has already died.

Table 1 presents a partial list of Hebrew and Aramaic key-words and key-phrases and a few acronyms in Hebrew and their translation into English.

4 Rule-Based Constraints

This section presents the rules, based on key-phrases and references, formulated for the estimation of the birth and death years of an author X (the extracted results point to specific years) based on his texts and the texts of other writers (Yi) who mention X or one of his texts. We assume that the birth years and death years of all writers are known, excluding those that are under interrogation. Now, we will give some notions and constants that are used: X – The writer under consideration, Yi – Other writers, B – Birth year, D – Death year, MIN – Minimal age (at present, 30 years) of a rabbinic writer when he starts to write his response, MAX – Maximal age (at present, 100 years) of a rabbinic author, and RABBI_DIS – The gap age between rabbi and his student (at present, 20 years). The estimations of MIN, MAX, and RABBI_DIS constants are heuristic, although they are realistic on the basis of typical responsa authors' lifestyles.

Different types of references exist: general references with and without key-phrases, such as "rabbi", "friend" and "late". There are two types of references: those referring

Table 1. Hebrew and Aramaic cue words – a partial list

Set	Key-phrase in Hebrew	Translation of the key-phrase
Late	זכור לטוב	Remembered for good
	זכרו לברכה קדוש צדיק וטהור	of blessed memory; holy, righteous and pure
	זכרו לברכה	of blessed memory
	זכר קדוש וצדיק לברכה	may the holy and righteous be of blessed memory
	זצוק"ל	acronym: may the righteous and holy be of blessed memory
	זקוצ"ל	acronym: may the holy and righteous be of blessed memory
	יז"ל	acronym: May his memory be forever
	ז"ל	acronym: of blessed memory
Friend	ידידי הרב הגדול	My friend the Great Rabbi
	ידידי הרב הגאון	My friend the Gaon Rabbi
	ידידי הרב	My friend (the) Rabbi
	השם ישמרהו ויחיהו	may G-D preserve him and grant him life
	ידידו הקטן	his young friend
	ישמרהו השם ויחיהו	may G-D preserve him and grant him life
	ידידי הרה"ג	partially acronym: My friend the Gaon Rabbi
	יה"ו	acronym: may G-D preserve him and grant him life
Rabbi	יורנו רבנו	May the Rabbi guide us
	מרא דאתרא	the local rabbinic authority
	רבי ומורי	My Rabbi and teacher
	מרנא	Our teacher
	רבינו	Our Rabbi

to living authors and those referring to dead authors. In contrast to academic papers, responsa include many more references to dead authors than to living authors.

We will introduce rules based on key-phrases and references of different degrees of certainty: "iron-clad" (I), heuristic (H) and greedy (G). "Iron-clad" rules are always true, without any exception. Heuristic rules are almost always true. Exceptions can occur because the heuristic estimates for MIN, MAX and RABBI_DIS are incorrect. Greedy rules are rather reasonable rules for responsa authors. However, wrong estimates can sometimes be drawn while using these rules. Each rule will be numbered and its degree of certainty (i.e., I, H, G) will be presented in brackets.

4.1 "Iron-Clad" and Heuristic Rules with Key-Phrases

First, we present two general heuristic rules, which are based on regular references (i.e., without any key-phrase), based on authors that cite X.

General rule based on authors that were mentioned by X

$$\mathrm{D(X)} > \, = \, \mathrm{MAX(B(Yi))} + \mathrm{MIN} \quad (1(\mathrm{H}))$$

X must have been alive when he referred to Yi, so we can use the earliest possible age of publishing of the latest born author Yi as a lower estimate for X's death year.

General rule based on authors that referred to X

$$\mathrm{B(X)} < \, = \, \mathrm{MIN(D(Yi))} - \mathrm{MIN} \quad (2(\mathrm{H}))$$

All Yi must have been alive when they referred to X, and X must have been old enough to publish. Hence, we can use the earliest death year amongst such authors Yi as an upper estimate of X's earliest possible publication age (and thus his birth year).

General rules based on year mentioning Y that appeared in X's documents

$$\mathrm{D(X)} > \, = \, \mathrm{MAX(Y)} \quad (3(\mathrm{I}))$$

X must have been alive when he mentioned the year Y. We can use the most recent year mentioned by X to evaluate the death year of X as an estimation of X's death year.

Posthumous key-phrase rules. Posthumous rules estimate the birth and death years of an author X based on references of authors who refer to X with the key-phrase "late" ("of blessed memory") or on references of X that mention other authors with the key-phrase "late". Figure 1 describes possible situations where various types of authors Yi (i = 1, 2, 3) refer to X with the key-phrase "late". The lines depict writers' life spans; the left edges represent the birth years and the right edges represent death years. In this case (as all Yi refer to X with the key-phrase "late"), we know that all Yi passed away after X, but we do not know when they were born in relation to X's birth. Y1 was born before X's birth; Y2 was born after X's birth but before X's death; and Y3 was born after X's death.

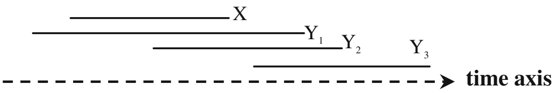

Fig. 1. References mentioning X with the key-phrase “late”.

$$D(X) < = \mathrm{MIN}(D(Yi)) \; (4(I))$$

However, we know that X must have been dead when Yi referred to him with the key-phrase “late”; thus, we can use the earliest born Y’s death year as an upper estimate for X’s death year. Like all writers, dead writers of course have to comply with rule (2) as well.

Now, we look at the cases where the author X that we are studying refers to other authors Yi with the key-phrase “late”. Figure 2 describes possible situations where X refers to various types of authors Yi (i = 1, 2, 3) with the key-phrase “late”. All Yi passed away before X’s death (or X may still be alive). Y1 died before X’s birth; Y2 was born before X’s birth and died when X was still alive; Y3 was born after X’s birth and passed away when X was still alive.

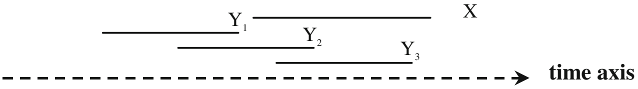

Fig. 2. References by X who mentions others with the key-phrase “late”.

$$D(X) > = \mathrm{MAX}(D(Yi)) \; (5(I))$$

X must have been alive after the death of all Yi who were referred by him with the key-phrase “late”. Therefore, we can use the death year of the latest-born Y as a lower estimate for X’s death year.

$$B(X) > = \mathrm{MAX}(D(Yi)) - \mathrm{MAX} \; (6(H))$$

X was probably born after the death year of the latest-dying person that X wrote about. Thus, we use the death year of the latest-born Y minus his max life-period as a lower estimate for X’s birth year.

Contemporary key-phrases rules. Contemporary key-phrases rules calculate the upper and lower bounds of the birth year of a writer X based only on the references of known writers who refer to X as their friend/rabbi. This means there must have been at least some period of time when both were alive together. Figure 3 shows possible situations where various types of writers Yi refer to X as their friend/rabbi. Y1 was born before X’s birth and died before X’s death; Y2 was born before X’s birth and died after X’s death; Y3 was born after X’s birth and passed away before X’s death; Y4 was

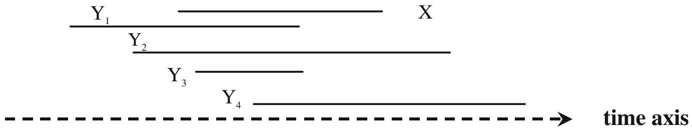

Fig. 3. References by authors who refer to X as their Friend/Rabbi.

born after X's birth and passed away after X's death. Like all writers, contemporary authors of course have to comply with rules 1 and 2 as well.

$$B(X) > = \mathrm{MIN}(B(Yi)) - (\mathrm{MAX} - \mathrm{MIN}) \; (7(H))$$

All Yi must have been alive when X was alive, and all of them must have been old enough to publish. Thus, X could not have been born MAX-MIN years before the earliest birth year amongst all authors Yi.

$$D(X) < = \mathrm{MAX}(D(Yi)) + (\mathrm{MAX} - \mathrm{MIN}) \; (8(H))$$

Again, all Yi must have been alive when X was alive, and all of them must have been old enough to publish. Hence, X could not have been alive MAX-MIN years after the latest death year amongst all writers Yi.

4.2 Greedy Rules

Greedy rules bounds are sensible but can sometimes lead to wrong estimates.

Greedy rule based on authors who are mentioned by X

$$B(X) > = \mathrm{MAX}(B(Yi)) - \mathrm{MIN} \; (9(G))$$

Many of the references in our research domain relate to dead authors. Thus, most of the references within X's texts relate to dead authors. Namely, many Yi were born before X's birth and died before X's death. Thus, a greedy assumption would be that X was born no earlier than the birth of the latest author mentioned by X; however, because there may be at least one case where Y was born after X was born, we subtract MIN.

Greedy rule based on references to year Y made by X

$$B(X) > = \mathrm{MAX}(Y) - \mathrm{MIN} \; (10\,(G))$$

When X mentions years, he usually writes the current year in which he wrote the document or a few years ahead. Most of the time, the maximum year, Y, minus MIN is larger than X's birth year.

Greedy rule based on authors who refer to X

$$D(X) <= \mathrm{MIN}(D(Yi)) - \mathrm{MIN} \quad (11(G))$$

As mentioned above, most of the references within Yi texts refer to X as being dead. Hence, most Yi died after X's death. Therefore, a greedy assumption would be that X died no later than the death of the earliest author who referred to X minus MIN.

Rules refinements 9–11 are presented by rules 12–17. Rules 12–14 are due to X referring to Yi and rules 15–17 are due to Yi referring to X.

Greedy rule for defining the birth year based only on authors who were referred to by X with the key-phrase "late"

$$B(X) >= \mathrm{MAX}(D(Yi)) - \mathrm{MIN} \quad (12(G))$$

When taking into account only references that were written in X's texts, most of the references are related to dead authors. That is, most Yi died before X's birth. Moreover, an author does not write from his birth; rather, he usually begins near his death. Thus, a greedy assumption would be that X was born no earlier than the death of the latest author mentioned by X minus MIN.

Greedy rule for defining the birth year based only on authors who are mentioned by X with the key-phrase "friend"

$$B(X) <= \mathrm{MIN}(B(Yi)) + \mathrm{RABBI\_DIS} \, (13(G))$$

When taking into account only references that are mentioned by X, which are related to contemporary authors, a greedy rule could be that X was born no later than the birth of the earliest author mentioned by X with the key-phrase "friend". Because many times the older author refers to the younger author as "friend", we need to add RABBI_DIS.

Greedy rule for defining the birth year based only on authors who are mentioned by X with the key-phrase "rabbi"

$$B(X) <= \mathrm{MIN}(B(Yi)) + \mathrm{RABBI\_DIS} \, (14(G))$$

When taking into account only references written in X's texts, which are related to contemporary authors, a greedy rule could be that X was born no later than the birth of the earliest author mentioned by X as a "rabbi". Due to the age difference between a student and his rabbi being approximately 20 years, we need to add RABBI_DIS.

Greedy rule for defining the death year of X based only on authors who referred to X with the key-phrase "late"

$$D(X) <= \mathrm{MIN}(B(Yi)) + \mathrm{MIN} \quad (15(G))$$

When taking into account only references written in Yi texts that refer to X with the key-phrase "late", a greedy assumption could be that X died no later than the birth of the earliest author who referred to X with the key-phrase "late"; because an author does not writes from birth, we need to add MIN.

Greedy rule for defining the death year of X based only on authors who referred to X with the key-phrase "friend"

$$D(X) > = MAX(D(Yi)) - RABBI\_DIS \quad (16(G))$$

When taking into account only references written in Yi texts that refer to X with the key-phrase "friend", all Yi must have been alive when X was alive, and all of them must have been old enough to publish; also, many times, the older author refers to the younger author with the key-phrase "friend", and the opposite never occurs. Therefore, a greedy assumption would be that X died no earlier than the death of the latest author who referred to X with the key-phrase "friend" minus RABBI_DIS.

Greedy rule for defining the death year of X based only on authors who referred to X with the key-phrase "rabbi"

$$D(X) > = MAX(D(Yi)) - RABBI\_DIS \quad (17(G))$$

This follows the same principle as the rule for defining the birth year, but because this time the student mentions the rabbi, we need to reduce RABBI_DIS.

4.3 Birth and Death Year Tuning

Application of the Heuristic and Greedy rules can lead to abnormalities, such as an author's death age being unreasonably old or young. Another possible anomaly is that the algorithm may result in a death year greater than the current year (i.e., 2015). Hence, we added some tuning rules: D – death year, B – birth year, age = D–B.

Current Year: if (D > 2015) {D = 2015}, i.e., if the current year is 2015, then the algorithm must not give a death year greater than 2015.

Age: if (age > 100), {z = age–100; D = D–z/2; B = B+z/2}, and if (age < 30), {z = 30 – age; D = D+z/2; B = B−z/2}. Our postulate is that a writer lived at least 30 years and no more than 100 years. Thus, if the age according to the algorithm is greater than 100, we take the difference between that age and 100, and then we divide that difference by 2 and normalize D and B to result in an age of 100.

5 The Model

The main steps of the model are presented below.

1. **Cleaning the texts.** Because the responsa may have undergone some editing, we must make sure to ignore the possible effects of differences in the texts resulting from variant editing practices. Therefore, we eliminate all orthographic variations.

2. **Boosting extracting key-phrases and key-words.**
3. **Normalizing the** references **in the texts.** For each author, we normalize all types of references that refer to him (e.g., various variants and spellings of his name, books, documents and their nicknames and abbreviations). For each author, we collect all references syntactic styles that refer to him and then replace them with a unique string.
4. **Building indexes**, e.g., authors, references to "late"/"friend"/"rabbi", and calculating the frequencies of each item.
5. **Performing various combinations of "iron-clad" and heuristic rules** on the one hand **and greedy rules** on the other hand **to estimate** the birth and death years of each tested author.
6. **Calculating averages** for the best "iron-clad", heuristic and greedy versions.

6 Examined Corpus, Experiments and Results

The documents of the examined corpus were downloaded from Bar-Ilan University's Responsa Project. The examined corpus contains 15,495 responsa written by 24 scholars, averaging 643 files for each scholar. The total number of characters in the whole corpus is 127,683,860 chars, and the average number of chars for each file is 8,240 chars. These authors lived over a period of 229 years (1786–2015). These files contain references; each reference pattern can be expanded into many other specific references [22].

Reference identification was performed by comparing each word to a list of 339 known authors and many of their books. This list of 25,801 specific references refers to the names, nicknames and abbreviations of these authors and their writings. Basic references were collected and all other references were produced from them.

We split the data into two corpora: (1) 10,512 responsa authored by 12 rabbis, with an average of 876 files for each scholar and each file containing an average of 1800 words spread over 135 years (1880–2015); (2) 15,495 responsa authored by 24 rabbis, with an average of 643 files for each rabbi and each file containing an average of 1609 words spread over 229 years (1786–2015) (the set of 24 rabbis contains the group of 12 rabbis). For more detailed information on the data set, refer to Table 2 in the appendix at the end of this article.

Because of the nature of the problem, it is difficult to appraise the results in the sense that although we can compare how close the system guess is to the actual birth or death years, we cannot assess how good the results are, i.e., there is no real notion of what a 'good' result is. For now, we use the notion Distance, which is defined as the estimated value minus the ground truth value.

The outcomes appear in the following histograms. Each histogram shows the results of one algorithm – Iron + Heuristic (I + H) or Greedy. Each algorithm was performed on two groups of authors: a group of 12 writers and a group of 24 writers. For both algorithm executions, there are outcomes containing estimated birth years and death years. The results shown in the histograms are the best birth/death date deviation results. In every histogram, there are eight columns; there are two quartets of columns in each histogram: the right quartet indicates the deviation from the death year, while the left quartet indicates deviations from the birth year. Each column represents a deviation without a key-phrase or with the year that was mentioned in the text, a deviation with the "late" key-phrase, with the "rabbi" key-phrase, and with the "friend" key-phrase. Moreover, we used two manipulations – Age and Current year. The column with a gray background contains the best results. Each histogram contains 8 columns (results); there are 16 histograms, so there are, in total, 128 results.

The Age manipulation is very helpful; we used it in 94.5 % of the experiments (i.e., 121/128 = 0.945) for all of the refinements, in both algorithms, with or without constants.

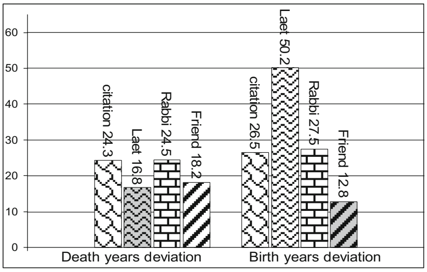

Fig. 4. 12 authors I + H no constant

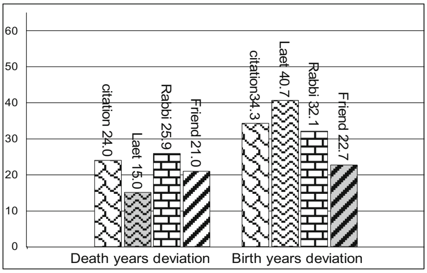

Fig. 5. 24 authors I + H no constant

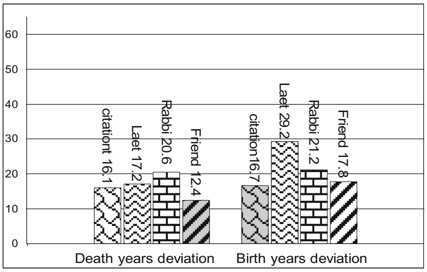

Fig. 6. 12 authors Greedy no constant

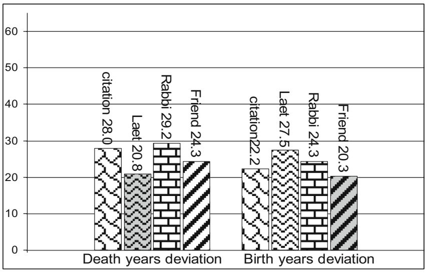

Fig. 7. 24 authors Greedy no constant

Examination of the effect of mentioning a year, listed in Figs. 4, 5, 6, and 7, compared with Figs. 8, 9, 10, and 11 regarding death year deviation, indicates that the contribution of referencing a year leads to an improvement of 2.8 years on average.

This phenomenon is more noticeable in Iron + Heuristic (average upswing of 4.2 years) than with Greedy (average deviation upswing of 1.4 years). The main reason for this is that a writer usually writes until close to his death. Additionally, when a year is mentioned in the text, it is often the year in which the writer wrote the document.

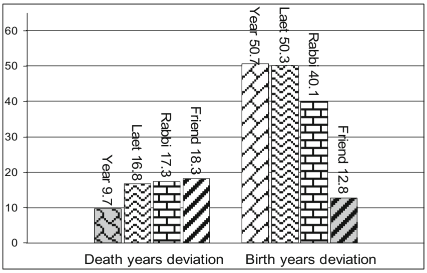

Fig. 8. 12 authors I + H no constant

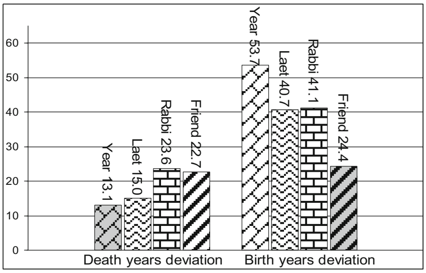

Fig. 9. 24 authors I + H no constant

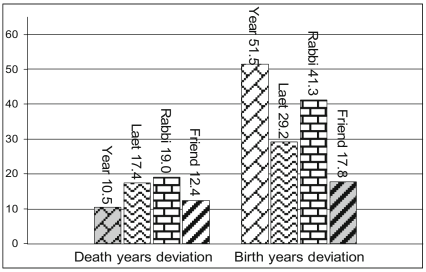

Fig. 10. 12 authors Greedy no constant

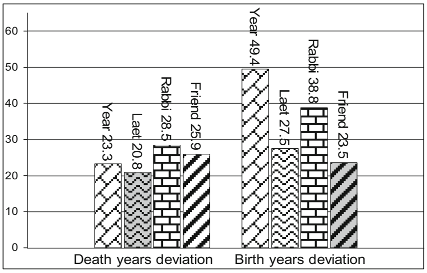

Fig. 11. 24 authors Greedy no constant

Because an author writes, in many cases, until near his death year, the maximum year mentioned in his texts is close to the year of his death.

In contrast to the death year assessment, birth year assessment has a negative impact; the deviation increases by 10.4 years, on average. It is essential to note that we are now evaluating the impact of the year mentioned in the text. If the results without using the year mentioned are better than the results using the year mentioned, it means that we should not use it. For example: the result of the birth year using Greedy rules, without year mentioning and without any refinement, for the 12 authors has a deviation of 16.7 years. After using the year mentioned, the deviation is 51.5 years, decreasing the accuracy by 34.83 years. The result of the birth year using the Iron + Heuristic, without year mentioning and without key-phrases, for the 12 authors has a deviation of 26.5 years. After using a reference to years, the deviation is 50.7 years, decreasing the accuracy by 24.2 years, i.e., the deviation with the use of year mentioning is greater. An analysis of the formulas shows that the formula that determines the birth year in the Greedy (10(G)) uses the most recent year the writer writes in his texts. The most recent year that the rabbi mentions is usually near his death, as explained above; therefore, very poor birth results are obtained, with a decline of 12.5 years. The results of the Greedy are better than Iron + Heuristic (decline of 8.4 years), but the effect of year mentioning on the results of Iron + Heuristic is less harmful. Thus, to estimate the death year, we will use the Iron + Heuristic algorithm with the use of year mentioning without any key-phrases.

The use of the key-phrase "friend" for birth year assessment gives the best results compared with the other key-phrases – "late", "rabbi" or none. This is because friends are of the same generation and more or less the same age; thus, they are born in roughly the same year. Thus, for a writer addressing another author as his friend, the assessment of his birth year will give good results. For the death year, however, this is not assured because there may be a much greater period between the deaths of friends (one may die at the age of 50, while his friend at the age 75). Hence, the "friend" key-phrase usually gives better birth year assessment than death year assessment (Figs. 12, 13, 14, and 15).

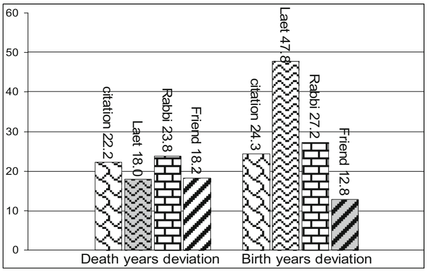

Fig. 12. 12 authors I + H with constant

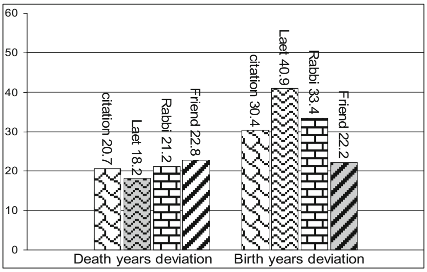

Fig. 13. 24 authors I + H with constant

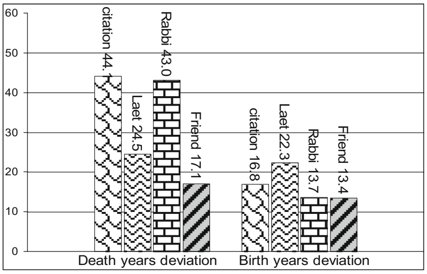

Fig. 14. 12 authors Greedy with constant

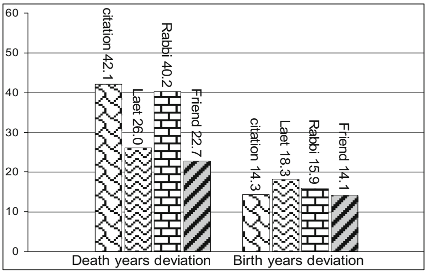

Fig. 15. 24 authors Greedy with constant

After we found that the best results for the birth year are always with the "friend" key-phrase (except for one case), we investigated at greater depth and found that this occurs specifically with the use of constants. Constants are important, resulting in an average improvement of 6.3 years in the case of Greedy (for the 12 and 24 authors). In general, a Posek is addressed in responsa after he has become important enough to be mentioned and regarded in the Halachic Responsa, which is usually at an advanced age.

We stated above that the use of Greedy rules with constants gives the greatest improvement. Even without the use of constants, Greedy produces the best results.

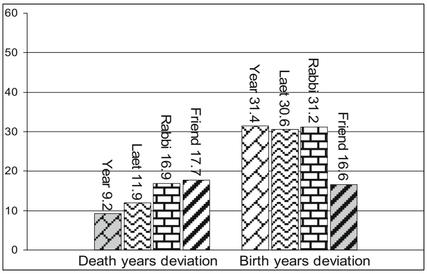

Fig. 16. 12 authors I + H with constant and year

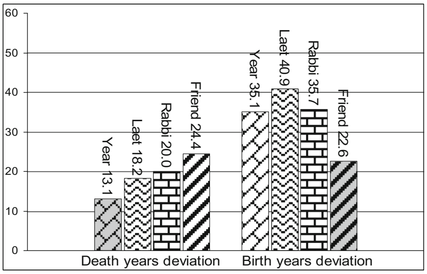

Fig. 17. 24 authors I + H with constant and year

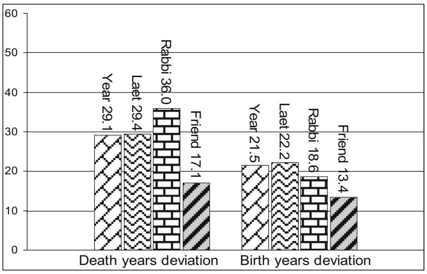

Fig. 18. 12 authors Greedy with constant and year

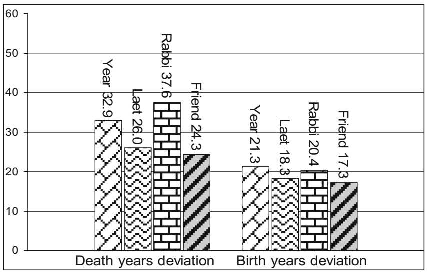

Fig. 19. 24 authors Greedy with constant and year

The reason lies in the formulae; formula (13(G)) finds the lowest birth year from the group of authors that the arbiter mentioned. Unlike the Greedy, the Iron + Heuristic formula (7(H)) reduces the constant (at present, 20); therefore, the results of the Greedy are better. In conclusion, to best assess the birth year, we apply the Greedy algorithm, using constants and also the key-phrase “friend”.

The best results when evaluating birth year occurred when using the Greedy algorithm with constants and without mentioning years. The best results when evaluating death year occurred when using the Iron + Heuristic algorithm with constants and without mentioning years. When we compare these results with the results shown in Figs. 16, 17, 18, and 19 we find that in the case of Greedy, when we add more authors, there is an improvement only in one case, i.e., for the 12 authors using the “late” key-phrase; the remaining results show a decline in performance. The reason for this phenomenon may lie in the Greedy formula; when an author is more successful, in addition to being mentioned by others many times, he is mentioned at a younger age by authors that are older than him; therefore, the estimation is less accurate. For example: the estimation of the death year of the late Rabbi Ovadia Yosef has an error of 61 years (instead of 2014, the algorithm result is 1953), determining that he died at an age of 34;

using the Iron + Heuristic algorithm, there was a decrease in two results and an improvement in 5 results. For Iron + Heuristic, there is an average improvement of 0.64 years and, in fact, the best death year result estimation. The quality of the Greedy algorithm birth year results estimation using year mentioning pretty severely impairs the results (explained above). A possible explanation for this is that the improvement that comes from using constants cannot overcome the deterioration that comes from year mentioning. In contrast, when assessing the death year, using year mentioning with Iron + Heuristic significantly improves the results, and using constants improves them a little more; therefore, a combination of constants + year mentioning brings better assessment of the death year. Therefore, when assessing birth year and death year, it is not enough to use references; we have to use key-words and key-phrases. To estimate death year, we will use the year(s) mentioned in the text and constants with the Iron + Heuristic algorithm; to estimate birth year, we will run the Greedy algorithm using constants and the "friend" key-phrase without year mentioning.

7 Summary, Conclusions and Future Work

We investigated the estimation of the birth and death years of authors using year mentioning, the "late" ("of blessed memory") key-phrase, the "rabbi" key-phrase, the "friend" key-phrase and undated references that are mentioned in documents of other dated authors that refer to author being considered or those mentioned by him. This research was performed on responsa documents, where special writing rules are applied. The estimation was based on the author's texts and texts of other authors who refer to the discussed author or are mentioned by him. To do so, we formulated various types of iron-clad, heuristic and greedy rules. The best birth year assessment was achieved by using the Greedy algorithm with constants and the "friend" key-phrase. The best death year assessment was achieved by using the Iron + Heuristic algorithm with year mentioning.

We plan to improve this research by (1) testing new combinations of iron-clad, heuristic and greedy rules, as well as a combination of key-phrases (e.g., "late" and "friend"); (2) improving existing rules and/or formulating new rules; (3) defining and applying heuristic rules that take into account various details included in the responsa, e.g., events, names of people, new concepts and collocations that can be dated; (4) conducting additional experiments using many more responsa written by more authors to improve the estimates; (5) checking why the iron-clad, heuristic and greedy rules tend to produce more positive differences; and (6) testing how much of an improvement we can obtain from a correction of the upper bound of D(x) and how much we will, at some point, use it for a corpus with long-dead authors.

Appendix

Data Set Information

Table 2. Full details about the data set

# of authors		Death year	Birth year	Author's name	# of files	# of words	# of characters
24	12	2015	1914	Vozner Shmuel	1807	1,490,463	7,768,059
		2014	1920	Yosef Ovadya	1283	4,578,049	22,933,473
		2006	1917	Waldenberg Eliezer	1639	3,197,662	16,589,888
		1995	1910	Auerbach Shlomo Zalman	229	793,706	4,087,592
		1989	1902	Weiss Yitzchak	1468	2,311,927	11,695,021
		1989	1911	Stern Bezalel	663	1,080,452	5,390,661
		1986	1895	Feinstein Moshe	1831	2,306,526	11,959,224
		1969	1890	Hadaya Ovadia	210	713,341	3,683,787
		1963	1898	Ades Yaakov	131	310,585	1,604,218
		1959	1901	Havita Rahamim	736	898,543	4,655,681
		1959	1889	Herzog Yitzchak	190	430,259	2,210,586
		1953	1880	Ben-Zion Meir Hai Uziel	374	899,617	4,621,414
		1948	1880	Boimel Yehoshua	129	237,093	1,227,007
		1942	1873	Baer Weiss Yitzchak	497	243,789	1,257,633
		1935	1865	Kook Abraham Yitzchak	681	750,145	3,892,610
		1921	1854	Allouch Faraji	112	205,258	1,069,460
		1911	1835	Schwadron Sholom Mordechai	1574	1,657,860	8,560,084
		1889	1813	Somekh Abdallah	86	80,508	412,486
		1896	1817	Spektor Yitzchak Elchanan	301	1,159,019	5,843,696
		1893	1820	Trunk Israel Yehoshua	281	132,257	689,598
		1880	1790	Abuhatzeira Yaakov	146	177,411	917,682
		1874	1801	Edery Abraham	119	176,849	918,564
		1866	1794	Assad Yehuda	882	880,361	4,565,230
		1843	1786	Birdugo Yaakov	126	218,402	1,130,206

References

1. Wintner, S.: Hebrew computational linguistics: past and future. Artif. Intell. Rev. **21**(2), 113–138 (2004)
2. HaCohen-Kerner, Y., Kass, A., Peretz, A.: HAADS: a Hebrew Aramaic abbreviation disambiguation system. J. Am. Soc. Inf. Sci. Technol. JASIST **61**(9), 1923–1932 (2010)
3. Gutwin, C., Paynter, G., Witten, I., Nevill-Manning, C., Frank, E.: Improving browsing in digital libraries with key-phrase indexes. Decis. Support Syst. **27**(1), 81–104 (1999)
4. Zhang, Y., Zincir-Heywood, N., Milios, E.: World wide web site summarization. Web Intell. Agent Syst. **2**(1), 39–53 (2004)
5. Hulth, A., Megyesi, B.B.: A study on automatically extracted key-words in text categorization. In: Proceedings of the 21st International Conference on Computational Linguistics and the 44th Annual Meeting of the ACL, pp. 537–544 (2006)
6. Kim, S.N., Baldwin, T.: Extracting key-words from multi-party live chats. In: Proceedings of the 26th Pacific Asia Conference on Language, Information, and Computation, pp. 199–208 (2012)
7. Berend, G.: Opinion expression mining by exploiting key-phrase extraction. In: IJCNLP, pp. 1162–1170 (2011)
8. Liu, Z., Huang, W., Zheng, Y., Sun, M.: Automatic key-phrase extraction via topic decomposition. In: Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing, ACL, pp. 366–376 (2010)
9. Hasan, K.S., Ng, V.: Conundrums in unsupervised key-phrase extraction: making sense of the state-of-the-art. In: Proceedings of the 23rd International Conference on Computational Linguistics: Posters, ACL, pp. 365–373 (2010)
10. Medelyan, O., Frank, E., Witten, I.H.: Human-competitive tagging using automatic key-phrase extraction. In: Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 3, ACL, pp. 1318–1327 (2009)
11. Kim, S.N., Medelyan, O., Kan, M.Y., Baldwin, T.: Automatic key-phrase extraction from scientific articles. Lang. Resour. Eval. **47**(3), 723–742 (2013)
12. Yih, W.T., Goodman, J., Carvalho, V.R.: Finding advertising key-words on web pages. In: Proceedings of the 15th International Conference on World Wide Web, pp. 213–222. ACM (2006)
13. Garfield, E.: Can citation indexing be automated? In: Stevens, M. (ed.) Statistical Association Methods for Mechanical Documentation, Symposium Proceedings, National Bureau of Standards Miscellaneous Publication 269, pp. 189–142 (1965)
14. Berkowitz, E., Elkhadiri, M.R.: Creation of a style independent intelligent autonomous citation indexer to support academic research, pp. 68–73 (2004)
15. Giuffrida, G., Shek, E.C., Yang, J.: Knowledge-based metadata extraction from PostScript files. In: Proceedings of the 5th ACM Conference on Digital libraries, pp. 77–84. ACM (2000)
16. Seymore, K., McCallum, A., Rosenfeld, R.: Learning hidden markov model structure for information extraction. In: AAAI-99 Workshop on Machine Learning for Information Extraction, pp. 37–42 (1999)
17. Ritchie, A., Robertson, S., Teufel, S.: Comparing citation contexts for information retrieval. In the 17th ACM Conference on Information and Knowledge Management (CIKM), pp. 213–222 (2008)
18. Bradshaw, S.: Reference directed indexing: redeeming relevance for subject search in citation indexes. In: Koch, T., Sølvberg, I.T. (eds.) ECDL 2003. LNCS, vol. 2769, pp. 499–510. Springer, Heidelberg (2003)

19. HaCohen-Kerner, Y., Beck, H., Yehudai, E., Rosenstein, M., Mughaz, D.: Cuisine: classification using stylistic feature sets and/or name-based feature sets. J. Am. Soc. Inf. Sci. Technol. (JASIST) **61**(8), 1644–1657 (2010)
20. HaCohen-Kerner, Y., Mughaz, D.: Estimating the birth and death years of authors of undated documents using undated citations. In: Loftsson, H., Rögnvaldsson, E., Helgadóttir, S. (eds.) IceTAL 2010. LNCS, vol. 6233, pp. 138–149. Springer, Heidelberg (2010)
21. Mughaz, D., HaCohen-Kerner, Y., Gabbay, D.: When text authors lived using undated citations. In: Lamas, D., Buitelaar, P. (eds.) IRFC 2014. LNCS, vol. 8849, pp. 82–95. Springer, Heidelberg (2014)
22. HaCohen-Kerner, Y., Schweitzer, N., Mughaz, D.: Automatically identifying citations in Hebrew-Aramaic documents. Cybern. Syst. Int. J. **42**(3), 180–197 (2011)