

THESIS TITLE

Jacob Timothy Hill

A dissertation submitted to the faculty at the University of North Carolina at Chapel Hill in partial fulfillment of the requirements for the degree of Doctor of Philosophy in the School of Information and Library Science.

Chapel Hill  
2020

Approved By:

Ryan Shaw

Melanie Feinberg

Jaime Arguello

Omid Ghaemmagami

Ted Underwood

©2020

Jacob Timothy Hill

ALL RIGHTS RESERVED

## **ABSTRACT**

# Contents

<b>1</b>	<b>Introduction</b>	<b>1</b>
<b>2</b>	<b>Literature Review</b>	<b>2</b>
<b>3</b>	<b>The Linguistic Setting</b>	<b>3</b>
3.1	The “Raw” Data . . . . .	3
3.2	Character Normalization . . . . .	5
3.3	A Range of Linguistic Modes . . . . .	9
3.4	Comparative Linguistics . . . . .	15
3.5	Conclusion . . . . .	16
<b>4</b>	<b>Word Embeddings</b>	<b>20</b>
<b>5</b>	<b>Conclusion</b>	<b>21</b>
	<b>Appendices</b>	<b>22</b>
	Appendix A: Bahá’í Published Volumes . . . . .	22

# List of Figures

3.1	Bahá'í works sampled relative to all extant works . . . . .	4
3.2	Venn diagrams depicting the effects of various processing decisions . . . . .	6
3.3	Venn diagram with Persian section highlighted . . . . .	13
3.4	Most common word lengths used by Bahá'u'llah . . . . .	16
3.5	Most common word lengths used by 'Abdu'l-Bahá . . . . .	17
3.6	Most common word lengths used by the Báb . . . . .	18
3.7	Most common word lengths used by al-Shaykh Murtaḍá al-Ánsarí . . . . .	19

# 1 Introduction

Overall argument: We are applying algorithms developed for information science and computer science tasks to humanities tasks and this is problematic because it requires significantly more interrogation of the process, the bias in the data (text encoding, archival bias e.g. what's included and not included, etc.).

When moving into a new field like Digital Humanities it is important to explore some questions related to methodological fit. These explorations have to be done for different fields and different problems. We can't have one solution for all humanities problems. The kind of work done thus far relates to some basic explorations of the application of algorithms to humanities problems, often without a clear idea of how to interpret the results, as well as some NLP work that addresses issues related to Arabic and Persian corpora. Before applying NLP problems to Bahá'í texts it is important to contextualize the problems which entails framing them in such a way that computer scientists and humanists, to adopt an overly simplistic schema of the audience, can engage in a common conversation.

## **2 Literature Review**

## 3 The Linguistic Setting

The Bahá'í writings are something of a linguistic anomaly.

Few writers have such a wide ranging audience in mind or adapt so much to the linguistic capacities/leanings of their audience.

### 3.1 The “Raw” Data

The sample of Bahá'í writings used in this study consists of nearly everything available on the authorized website of the Universal House of Justice ([www.reference.bahai.org](http://www.reference.bahai.org))—the elected governing body of the international Bahá'í community—with the exception of some early scholarly attempts to compile the Bahá'í writings, such as those of ‘Abdu’l-Hamíd Ishráq-Khávári, and a few largely derivative compilations. Including the former would require significant manual labor to separate the words of the author from the passages of Bahá'í scripture he compiles. Including the latter would add duplicate text, possibly skewing the results of the study through the multiplication of commonly occurring passages. Volumes consisting of scanned images without machine readable text were naturally excluded because of the significant work required to prepare them for computation. For a list of all published volumes included in the present study see Appendix A: Bahá'í Published Volumes. For a relational view of the number of works in the sample compared to all extant works see Figure 3.1.

The individual Bahá'í works were parsed from the published volumes listed in Appendix A: Bahá'í Published Volumes and labelled with author and language (Arabic or Persian). This binary linguistic labelling at the level of the work may be somewhat misleading given the fluidity of language that is characteristic of the Bahá'í writings. A more granular labelling strategy—say at the level of the sentence or paragraph—though ideal, would require an exponential increase in manual



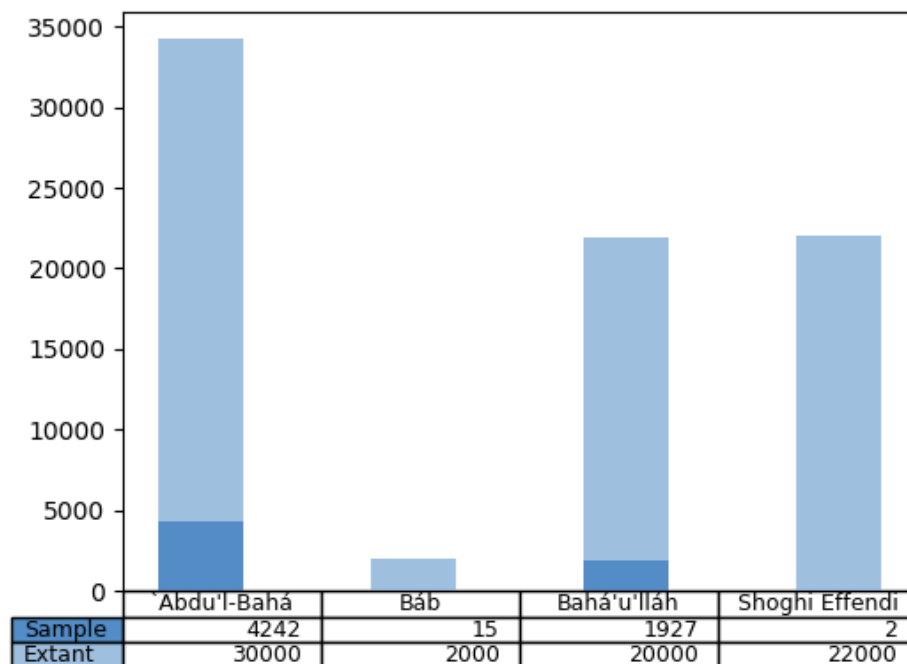


Figure 3.1: Bahá'í works sampled relative to all extant works

labor. By way of compromise, the works labelled 'Arabic' are fully Arabic. The occasional Persian word, or Persianized Arabic word, may slip in but the general structure is Arabic. There are no Persian sentences, prepositions, or pronouns. Works were labelled manually, checked computationally, reinspected and relabelled if necessary. The checking step consisted of generating a list of high frequency Persian words, which included all prepositions and pronouns, and iteratively searching each Arabic text for words in this list. If the Arabic text contained any of the high frequency Persian words, it was reinspected and, if necessary, relabelled.

This labelling strategy is justified by historical and cultural precedent. Persians, being largely Muslim, were expected to have some knowledge of the Arabic language—the language of the Qur'án. Arab Muslims, on the other hand, were not expected to learn Persian. The impact of this cultural and historical phenomenon on the Bahá'í corpus is discernible in the larger number of wholly Arabic works—particularly among earlier among the earlier authors—as well as the fre-

quency of lengthy Arabic passages in most Persian works. There are a significant number of Persian works without Arabic passages, but given the expectation that Persian readers would know some Arabic, the appearance of large segments of Arabic text in Persian works is commonplace.

The implications of such a strategy are that the total number of Arabic and Persian words are skewed towards Persian (there are 706,426 Arabic words compared to 1,906,607 Persian). Most of the “Persian” works contain large sections of Arabic text which gets counted in the Persian word count. This, unfortunately, is unavoidable given that a more granular labelling approach is not presently feasible.

In order to provide context, a large corpus of contemporary texts were selected for comparison, which included a collection of over 60,000 Persian poems (“and a few prose texts”) prepared by the Persian Digital Library Pilot Project (PDL) [1] and nearly 350 lengthy Arabic works from the ‘Knowledge, Information Technology and the Arabic Book’ (KITAB) [2]. The selections from the KITAB project were limited to works from the 12th-14th centuries AH (October 26, 1688-November 21, 1979) which is roughly contemporaneous with the Bahá’í works sampled. All works in the PDL corpus were added without exception while works in the KITAB corpus that contained Persian passages were removed to maintain consistency with the linguistic labelling strategy applied to the Bahá’í corpus.

## 3.2 Character Normalization

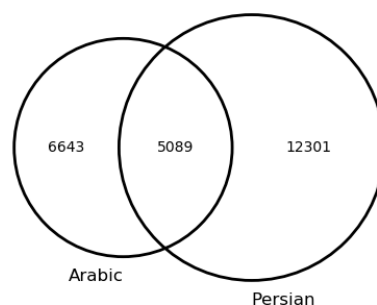
All computational linguistic work requires some level of pre-processing. There are several considerations that are peculiar to languages written in a modified form of the Arabic script. One such problem is the existence of ambiguous characters—characters with the same form, but different encodings [3]. These characters appear the same to the human eye but have different underlying encodings. What appears to be the same letter in Persian and Arabic—the letter *kāf* for example—may in fact be two distinct encodings. Failure to conflate ambiguous characters results in the proliferation of vocabulary; a significant number of words that appear the same to the reader and

were intended to be the same by the writer are treated as different words by the computer. Failure to remediate this issue prior to any down stream tasks can have significant negative effects on the outcome of those tasks.

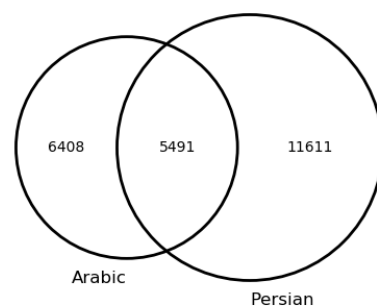
These effects can be especially consequential if the object of interest is the interactions across language as the shared vocabulary—the loan words shared between languages that evince centuries of linguistic and cultural exchange—are obliterated by the isolated processes through which the languages were encoded.

Another issue affecting languages written in Arabic script is the inconsistent use of diacritical marks. In Arabic and Persian short vowels were historically not written, though they are pronounced in speech. The reader is expected to know when to add the vowels, and which ones to add. There are, of course, a significant number of homographs which led to the subsequent development of diacritical marks that could be optionally added above or below letters to specify the following short vowel and thereby remove any ambiguity. Adoption has not been consistent, however, which compounds the vocabulary proliferation problem described above.

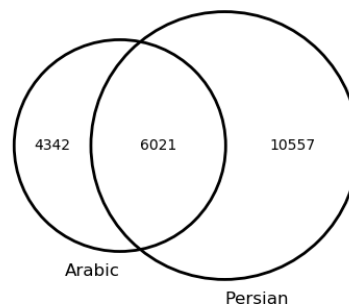
A number of steps were taken to remediate this issue, the effects of which are summa-



(a)



(b)



(c)

Figure 3.2: Arabic and Persian vocabulary with various levels of processing

rized in Figure 3.2. For all steps, the minimum frequency threshold was set at 5; words that occur less than five times across the whole corpus were removed. Diagram (a) shows the respective vocabulary with no data cleaning. Diagram (b) shows results after several language-agnostic cleanup steps were taken. Punctuation marks were removed as well as characters that produce only visual effects, such as spaces, tabs, and newline characters. Diagram (c) shows the results of removing character disambiguation and diacritical marks in addition to the steps taken in (b).

The term “pre-processing”, commonly used to describe this kind of work gives some indication of the level of importance it is given in the literature. Though it is deemed essential to any downstream task, in most computational linguistic research it is briefly described, or not at all. The field trends towards standardizing these practices and moving on to other, more interesting, things. Precedent often trumps rationale in determining which data cleaning steps are taken, and with good reason. Precedent has worked fine for decades and the circumstances have not changed. The application of computational linguistics strategies to humanities problems—problems they were not overtly designed to solve—warrants a re-examination and justification of these steps, as does the application to non-western languages. Both of these contextual changes introduce opportunities for confusion.

The Venn diagrams in Figure 3.2 provide insight into the effects pre-processing decisions can have. As is expected, without any intervention (a) there is a significant overlap in vocabulary. After all, many of the Persian texts contain Arabic passages and Persian, in general, is riddled with Arabic vocabulary. Language agnostic processing steps reduced the total vocabulary by 2.2% (523 words), the Arabic vocabulary by 3.5% (235 words), and the Persian vocabulary by 5.6% (690 words), while the intersection increased by 7.9% (402 words). Removing diacritical marks and conflating ambiguous characters reduced the total vocabulary by an additional 11% (2,590 words), the Arabic vocabulary by 32% (2,066 words), and the Persian vocabulary by 9.1% (1,045 words), while increasing the intersecting vocabulary by 9.7% (530 words). The reduction in vocabulary consisted exclusively of non-existent words—side effects of the text encoding process or

typographical errors.

It is worth noting, also, the effect the minimum frequency threshold has on vocabulary. One of the implications of Zipf's law on natural language is the common existence of a long tail of words that occur only once, or a few times. Removing these words helps to speed up computation and reduce noise; words that occur only once can have no interaction effects and words that occur only a few times do not provide enough data to draw statistically significant conclusions. With no pre-processing 243,639 words were removed, with language agnostic pre-processing only 163,692 words were removed, and with language agnostic processing, removal of diacritical marks and conflation of ambiguous characters only 103,824 words were removed. The trend towards discarding fewer words with each step of processing is a clear indication of the efficacy of these techniques; nearly 140,000 words were spared from removal. The raw count is not an adequate appraisal of the value of these techniques, however. Bearing in mind the objective of representing this vocabulary as a single matrix, these 140,000 words take on added significance.

To illustrate this point and simultaneously delineate the limitations of the adopted strategies consider the word 'kalima', an Arabic loan word meanings 'word'. The first letter 'kaf' has a different encoding in Arabic and Persian systems. Given that text encoding is not well understood by your average typist and considering that many of these typists may know Arabic and Persian and use both keyboards the word will bifurcate in both languages resulting in two forms in each language. The word 'kalima' has a further problem with its final letter. In the Arabic original form it ends in ta marbuta, a letter indicating the feminine gender of the word. This letter is not found in Persian and is replaced by the letter 'heh', a somewhat similar looking letter that also exists in Arabic. Adding in this results in four different possible forms for the same word—Arabic kaf... ta marbuta (form 1), Persian kaf... heh (form 2), Arabic kaf... heh (form 3), Persian kaf... ta marbuta (form 4). Normalizing the kaf would eliminate forms 2 and 4. For many words that completely remove the problematic effects of divergent text encodings. Unfortunately there is no easy solution for the heh ending adopted in Persian. Since it is used in Arabic, and often at the end of words, it cannot be blindly replaced with a ta marbuta. This would fix words like kalima

while creating significant noise in the process. Normalizing variant text encodings is only a solution for vocabulary proliferation problems resulting from the historical circumstances of text encoding in Arabic and Persian. It cannot solve problems resulting from other historical processes.

Failure to adequately prepare language data for down stream tasks delegitimizes the outcomes of those tasks. Meaningless sequences of encodings will erroneously be treated as words and examined for statistical significance while meaningful words will be discarded. This point is well established. What is not well established is the need for local processing models. Global language models-one size fits all approaches-can be problematic in unexpected ways. Languages have developed through unique historical and cultural circumstances. Text encoding decisions are a significant part of this historical process. Failure to adequately understand and respond to the historical processes in which the language developed and was encoded for computational processing can undermine the whole process.

### **3.3 A Range of Linguistic Modes**

The Bahá'í writings in general, and the writings of Bahá'u'llah in particular, are comprised of numerous styles or genres. Bahá'u'llah demonstrates an unusual, perhaps unique, ability to adjust the style of His writing to a multitude of ends. In a time and place where religious and cultural barriers seemed firmly fixed, He showed an uncanny ability to sweep them aside, making room thereby for a more embracing vision. In the religious and cultural diversity of His audience, in His demonstrated ability to write in a manner approachable to that audience, in His ability to transcend fixed literary genres, in the range of His vocabulary and grammatical prowess-in all these Bahá'u'llah shows a remarkable awareness of contemporary cultural, religious, and linguistic norms and a desire to transcend them in favor of ever wider horizons.

In the Suratu'l-Haykal, Bahá'u'llah wrote "We have revealed Our verses in nine different modes... Should it be Our wish, We would reveal them in countless other modes." A clear identification of these modes was never made by Bahá'u'llah or either of His authorized interpreters

and it is likely that He did not wish His audience to become overly fixated on such a classification schema. Jináb-i-Fádíl-i-Mázindarání, a prominent early scholar of the Bahá'í Faith, made a preliminary classification which attempted to identify these nine modes as those tablets with the tone of authority, those with the tone of servitude, those interpreting past scripture, those specifying of laws and ordinances, mystical writings, writings about government and world order, writings about the various branches of learning, writings calling for education, good character and virtues, and works treating social teachings [].

One limitation of Mázindarání's schema is that it constitutes a significant departure from Bábí and Islamic discourse without sufficient justification. The Arabic term *sha'n*, meaning mode or style, can be found in the writings of the Báb and, more generally, in Islamic thought. Nader Saiedi points out that there were four modes: divine verses; prayers and supplications; commentaries and sermons; and rational, educational, and philosophical discourse [4]. To these five modes, the Báb adds the Persian mode-the idea that God would speak in a language other than Arabic being quite revolutionary.

The four modes of Islam roughly correlate with the perspective of the speaker. The first mode (divine verses) represents the voice of God speaking directly to His creation. The second mode (prayers and supplications) is the voice of creation responding to its Creator. Together the two modes constitute a dialogue between the Creator and the creation. Linguistically, these modes would be distinguished from others by the prominence of first and second person pronouns and verbs constituted in the first and second person. See, for example, the two following passages, the first of which is in the mode of divine verses and the second in the mode of supplication and prayer: "O SON OF MAN! *I loved thy creation, hence I created thee. Wherefore, do thou love Me, that I may name thy name and fill thy soul with the spirit of life* [5]." "Glory be to *Thee, O my God! Thou hearest Thine ardent lovers lamenting in their separation from Thee, and such as have recognized Thee* wailing because of their remoteness from *Thy* presence [].".

The third (commentaries and sermons) and fourth mode (rational, educational, and philosophical discourse) would be characterized by a predominance of third person pronouns and verbs as

	<b>File Path</b>	<b>Word Count</b>	<b>Percent Persian</b>
1	bahauallah/text/bahauallah-amtm-21-fa.txt	323	79.88
2	bahauallah/text/bahauallah-amtm-22-fa.txt	255	91
3	bahauallah/text/bahauallah-amtm-23-fa.txt	182	91.76
4	bahauallah/text/bahauallah-amtm-24-fa.txt	187	76.5
5	bahauallah/text/bahauallah-amtm-25-fa.txt	522	82.76
6	bahauallah/text/bahauallah-amtm-37-fa.txt	252	71.43
7	bahauallah/text/bahauallah-aqa2-70-fa.txt	1292	78.49
8	bahauallah/text/bahauallah-st-004-fa.txt	130	86.15
9	bahauallah/text/bahauallah-st-058-2-fa.txt	206	72.82
10	bahauallah/text/bahauallah-st-098-fa.txt	165	72.12
11	bahauallah/text/bahauallah-st-199-fa.txt	45	75.6

Table 3.1: Works of Bahá'u'llah possibly written in a pure Persian

the subject matter is further removed from the speaker. Thus the four modes correspond to varying degrees of remoteness from the Divine creator and are affirmed, respectively, by the verses “I am God”, “Thou art God”, “He is God”, and “He is God Who is” [4]. To these modes, the Báb added the Persian mode, the content of which could be any of the other four modes [4].

It is clear that many of Bahá'u'llah's readers would have been familiar with this discourse and His statement regarding the revelation of His verses in “nine modes” would have conjured up this discourse in their minds. The first three categories in Mázindarání's classification overlap with those established in Islamic thought and reaffirmed by the Báb. Given that categories four through nine depart from the established norm they, in my mind, require more justification.

The point of such a departure is to demonstrate Bahá'u'llah's intentionality in writing in many different styles even if He never intended to, or we cannot concretely, crystallize His stylistic schema. Such a point is important given that the variety of styles-whether style is operationalized as subject matter, perspective, or something else entirely-are likely to carry linguistic traces that are discernable to computational processes and may have implications for the application or algorithms and the interpretation of their outputs.

Evidence of Bahá'u'llah's linguistic capacity can be found in the range of the vocabulary He uses in different works. The spectrum ranges from pure Persian works-those with no Arabic vocabulary-to pure Arabic works and everything in between. There was a movement contempo-



rary to Bahá'u'llah, and still in existence today, that aimed to cleanse the Persian language of all of its acquired Arabic baggage. Bahá'u'llah was not an advocate of this movement. In a moment of unusual candor He asserted: "There are many things which have not been expressed in Persian, that is to say, words referring to such things have not been devised, whilst in Arabic there are several words describing the same thing. Indeed there existeth no language in the world as vast and comprehensive as Arabic [6]." Yet, at times He deferred to the wonts of His correspondents and demonstrated His capacity to write in pure Persian. A query for works largely situated in the Persian sphere of the Venn diagram depicted in Figure 3.3 returns several works written in this style. If the cutoff level is set at 70%, that is at least 70% of the words present in the work are only present in other Persian works, 11 works are returned 3.1 which can roughly be clustered into three groups: false positives, intentionally pure Persian works, and works with relatively few Arabic words that are symptomatic of another cause.

Numbers 6, 9 and 11 are likely false positives. The fact that they have relatively few Arabic words is likely attributable to something other than the Bahá'u'llah's intention to write without the use of Arabic vocabulary. Work 11, for example, has only 45 words. The fact that such a small percentage of them are of Arabic origin is likely due to the sample size. Works 6 and 9 have a higher percentage of Arabic loan words than most of the works in this list and there seems to be no discernable pattern to their placement within the works.

Works 1, 4, 7, 8, 10 all have relatively few Arabic loan words though this is likely attributable to other factors. Works 1 and 4 are published in consecutive order with other tablets addressed to believers of Zoroastrian background [7]. The sparsity of Arabic vocabulary may stem from the Author's awareness of the cultural and linguistic background of His audience without the overt intention to avoid Arabic words altogether. Likewise, works 7, 8, and 10 share a genre. They are all poems reminiscent of the major Persian Sufi classical poets: 'Attár, Sa'di, Rumi and Háfez. A predominate Persian vocabulary sprinkled with Arabic words is characteristic of the genre.

There seems to be more deliberateness to the avoidance of Arabic words in works 2, 3, and 5. They were published in consecutive order with works 1 and 4 and were likewise addressed

to believers of Zoroastrian heritage. Works 1, 4, and 5 were later translated in a volume of other works addressed to early Zoroastrian converts to the Bahá'í Faith [6]. The introduction to the volume explicitly mentions that the recipient of work number 1, Mánikchí Šāhib, requested that Bahá'u'llah respond—and write exclusively henceforth—in pure Persian. Bahá'u'llah acknowledged this request by penning His reply in pure Persian, save the 75 word closing section which is a mixture of Arabic and Persian. Work number 2 found 9% of the words in the Arabic vocabulary but these words consist largely of homonyms such as the word 'shahr', which means 'announcement' or 'declaration' in Arabic and 'city' in Persian. The other words are clearly of Persian origin; their inclusion in the intersection of the Venn diagram Figure 3.3 must be attributed to their appearance in an Arabic work or, possibly, to a mislabelling of one or more files. Work number 3 is of the same type; all words except the word 'ism' are of Persian origin.

The examples above testify to Bahá'u'llah's ability to avoid words of Arabic origin altogether but they seem exceptional from a linguistic perspective. Generally His Persian writing is riddled with Arabic vocabulary. In many cases there are significantly more Arabic than Persian words. For example, in a letter to Zaynu'l-Muqarrabín, nearly eighty percent of the words are of Arabic origin [8].

There are many other such examples. In these cases the Persian language is like the mortar holding together the Arabic bricks; the verbs and some of the conjunctions are the only non-Arabic words in the text.

The same search strategy cannot be employed for revealing pure Arabic works. The fact that Arabic words are commonplace in the Persian language implies that the vocabulary of many pure Arabic works would have one foot firmly in the intersection between the two languages. Roughly

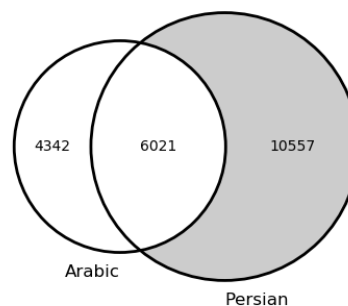


Figure 3.3: Pure Persian works can be identified by querying for works with vocabulary situated largely in the Persian section

60% of the Arabic vocabulary also appears in Persian works whereas only 36% of the Persian vocabulary is used in Arabic works. Given that the Persian words do not abound in the Arabic language, there are sufficient examples of Bahá'u'llah's Arabic with little to no Persian influence.

The quality of Bahá'u'llah's writing in both Arabic and Persian is something of a controversy. On one hand His writings have won the praise of renowned experts such as Khalil Gibran, according to a the third hand account of Marzieh Gail, [9] and Edward Granville Browne who wrote of the *Kitab-i-Iqan*: "it is a work of great merit, vigorous in style, clear in argument, cogent in proof, and displaying no slight knowledge of the Bible, Qur'an, and Tradition" [10]. On the other hand, He has amassed a large number of critics. Such opinions-both positive and negative-are highly subjective and of little use in determining Bahá'u'llah's linguistic competence. There are better ways to document Bahá'u'llah's linguistic versatility and prowess.

Style, or literary genre, is certainly a fluid category. Assigning boundaries to genres can be problematic. Yet identifying stylistic similarities between groups of literary works can shed light on questions of influence and awareness. The range of styles found among Bahá'u'llah's extant writings is noteworthy in this respect. Frank Lewis lists at least six distinct styles exemplified in His writings: the tradition of rhymed prose (*saj'*); *tafsir*; the classical Persian Sufi literary tradition exemplified by 'Attár, Sa'di, Rumi and Háfez; the *báz gasht* adabí style which began as a counter movement to the *sabk-eí* style; the gnomic tradition (*andarz*); and the Persian gnostic tradition [11]. William McCants notes the Shi'í narrative structure "underlying most of His epistles and homilies [12]." With little effort exemplars of other literary genres could be found in His writings. However, those already listed are sufficient to demonstrate the point that a number of disparate styles are contained within the corpus of Bahá'u'llah's writings.

An even more impressive example of Bahá'u'llah's command of Arabic and Persian and His ability to write with precision can be found in the *Hidden Words* [5]. At the time of the revelation of the *Hidden Words* (1857) Bahá'u'llah's audience would have been divided into four exclusive groups: Sunni, Shi'i, Wujudies and Shuhudis. Each of these, in turn, would be divided into opposing factions: Akhbaris, Usulis, and Shaykhis [13]. Bahá'u'llah wrote the *Hidden Words* in

such a way that it would appeal to each faction. Quotations are woven throughout in paraphrase, there is no mention of a proper name that could signal an allegiance to any faction, no isnads, no legalistic doctrines or cultic pronouncements [13]. What remains is something that could be seen by all as pure religion. “A religion apparently unencumbered by the tragedy of history, appearing as a restatement of basic truths through the medium of a compelling religious literary art in both languages of the city: Arabic (71 ‘verses’ and Persian (82 ‘verses’) [13].” Such an accomplishment is extraordinary given the sectarian nature of the religious discourse of 19th century Baghdad. Moreover, it shows a profound awareness of the discourse of each faction; in order to know which words to avoid, Bahá’u’llah had to be aware of which were most strongly identified with each sect. The work suggests a clear intent to elevate the minds of His audience beyond such sectarian discourse and serves as a profound example of His command of both languages.

### **3.4 Comparative Linguistics**

Another way to explore the linguistic characteristics of Bahá’u’llah’s writings is through comparison to other authors. There are a number of methods that have been used for this purpose, many of which cannot be applied to the writings of Bahá’u’llah. For example, Justin Rice uses average sentence length as a metric to distinguish the style of Hemmingway’s writings from that of other writers [14]. In classical Arabic and Persian, and in the writings of Bahá’u’llah specifically, the period is an afterthought. It is used inconsistently when it is used at all and it must, until proven otherwise for each text, be treated as what Gennete refers to a ‘paratextuality’—an element outside of the text that influences its reading [17]. Periods in English, and often in modern Arabic and Persian, are part of the author’s literary production; they are added by the author and can be used as evidence of his or her writing style. In the writings of Bahá’u’llah periods were added by scribes or publishers, if at all, and cannot therefore be used as evidence of writing style. Without periods there is no reliable way of determining sentence length which makes this metric inappropriate for the task at hand. Other examples could be given but this should be sufficient to demon-

strate that stylometric methods are context dependent; any given method cannot be assumed to hold accross language and time.

Among the methods that can safely be used to compare the writings of Bahá'u'llah to contemporary authors are average word length [14], lexical richness [14], and the chi-squared statistic [15] [16]. Word length

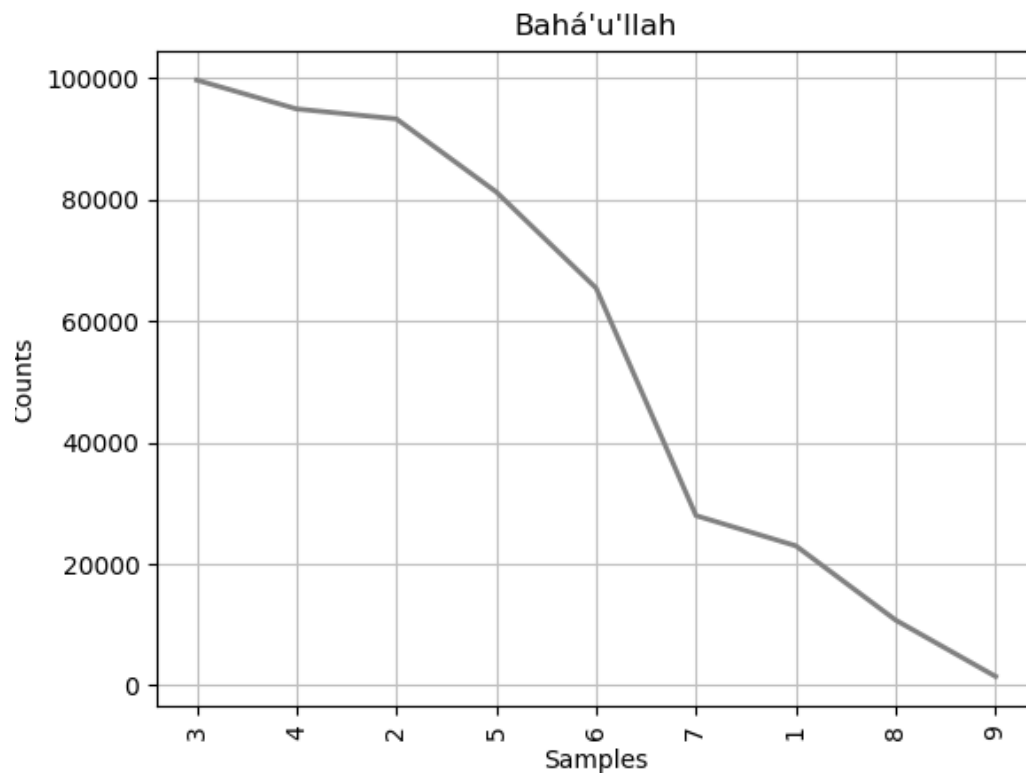


Figure 3.4: Most common word lengths used by Bahá'u'llah

### 3.5 Conclusion

Bahá'u'llah clearly adapted His writing style-genre, subject matter, and vocabulary-to His intended audience.

What you have is a highly deliberate collection of writings with ties to a plethora of genres and styles and yet linguistically distinct in many respects. That is, the position Bahá'u'lláh takes

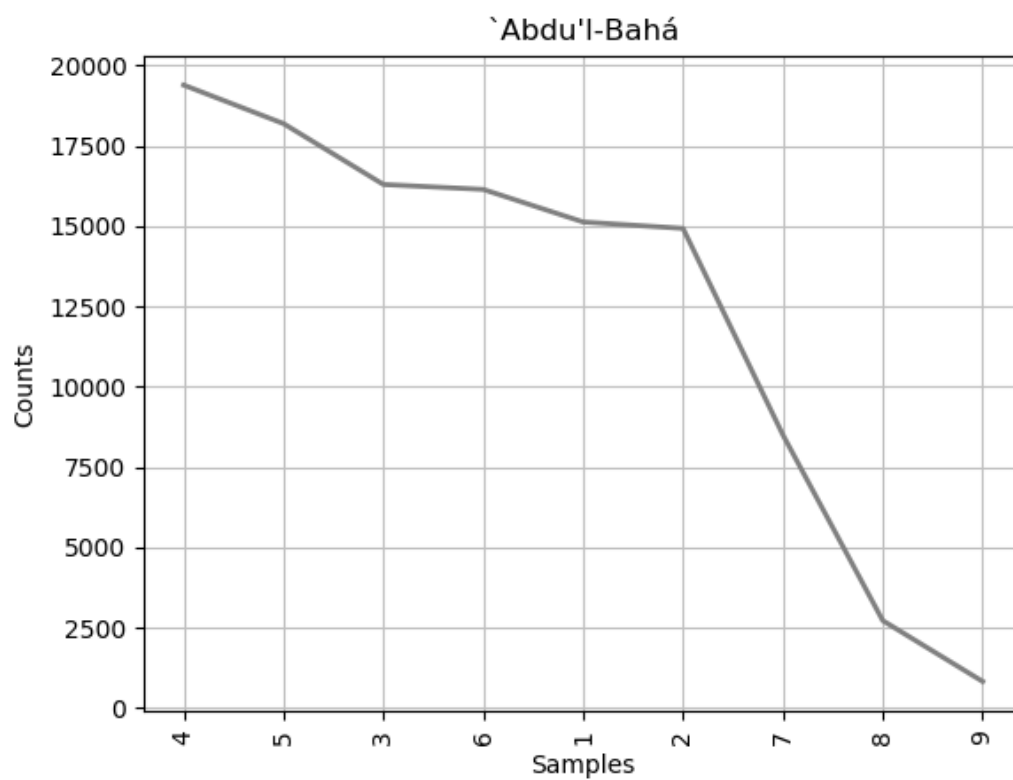


Figure 3.5: Most common word lengths used by ‘Abdu’l-Bahá

is both deliberately linked to past discourse and simultaneously distinct.

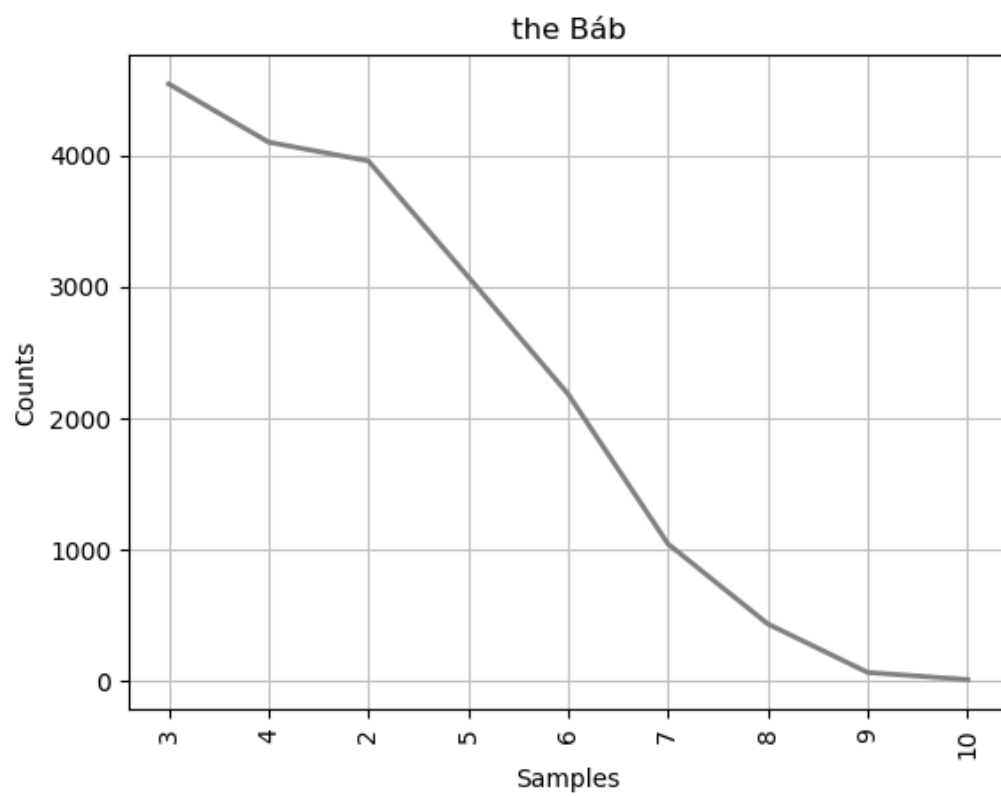


Figure 3.6: Most common word lengths used by the Báb

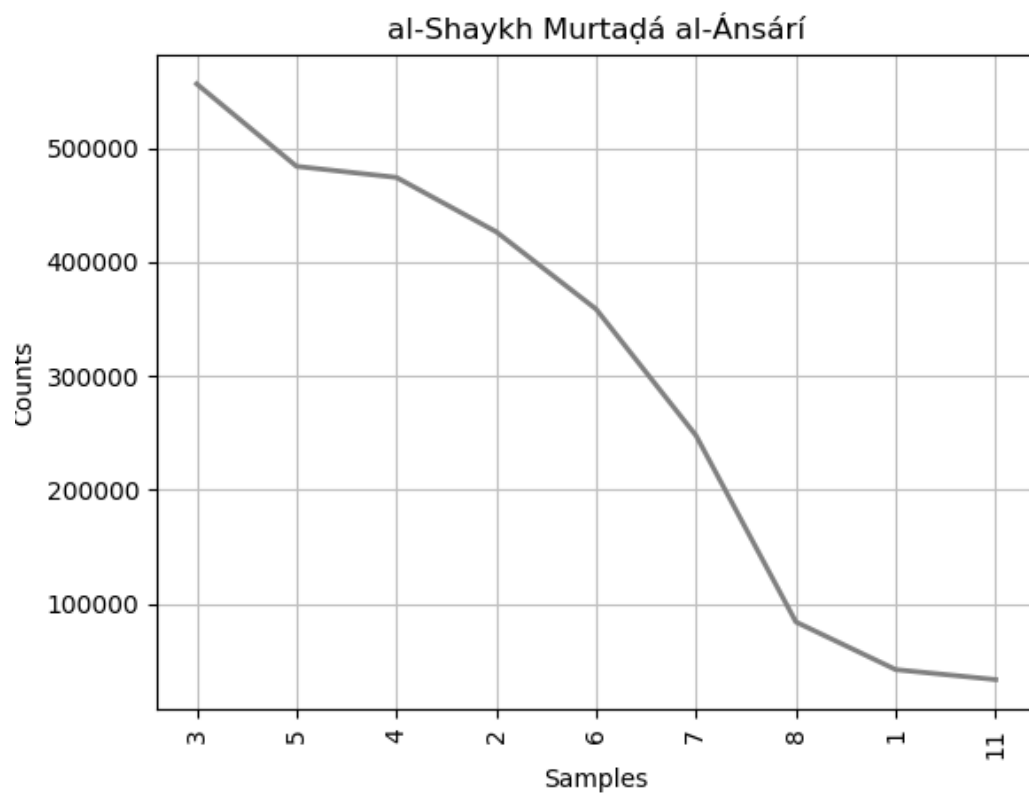


Figure 3.7: Most common word lengths used by al-Shaykh Murtaḍá al-Ānsarī



## **4 Word Embeddings**

## **5 Conclusion**

## Appendix A: Bahá'í Published Volumes

Author	Title
'Abdu'l-Bahá	Majmù'eh Monàjathà Haḍrat 'Abdu'l-Bahá
'Abdu'l-Bahá	Makatib Haḍrat 'Abdu'l-Bahá Jild 1
'Abdu'l-Bahá	Makatib Haḍrat 'Abdu'l-Bahá Jild 2
'Abdu'l-Bahá	Makatib Haḍrat 'Abdu'l-Bahá Jild 3
'Abdu'l-Bahá	Makatib Haḍrat 'Abdu'l-Bahá Jild 4
'Abdu'l-Bahá	Makatib Haḍrat 'Abdu'l-Bahá Jild 5
'Abdu'l-Bahá	Makatib Haḍrat 'Abdu'l-Bahá Jild 6
'Abdu'l-Bahá	Makatib Haḍrat 'Abdu'l-Bahá Jild 7
'Abdu'l-Bahá	Makatib Haḍrat 'Abdu'l-Bahá Jild 8
'Abdu'l-Bahá	Maqàleh Shakh̲si Síaḥ
'Abdu'l-Bahá	Mufāwadhāt
'Abdu'l-Bahá	Muntakhibāti az Makatib Haḍrat 'Abdu'l-Bahá Jild 1
'Abdu'l-Bahá	Muntakhibāti az Makatib Haḍrat 'Abdu'l-Bahá Jild 2
'Abdu'l-Bahá	Muntakhibāti az Makatib Haḍrat 'Abdu'l-Bahá Jild 3
'Abdu'l-Bahá	Muntakhibāti az Makatib Haḍrat 'Abdu'l-Bahá Jild 4
'Abdu'l-Bahá	Muntakhibāti az Makatib Haḍrat 'Abdu'l-Bahá Jild 5
'Abdu'l-Bahá	Muntakhibāti az Makatib Haḍrat 'Abdu'l-Bahá Jild 6
'Abdu'l-Bahá	Tadhkirat al-Vafā'
Báb	Muntakhibát Áyát az Áthár Haḍrat Nuḡteh Awlā
Bahá'u'lláh	Kitáb-i-Àqdas
Bahá'u'lláh	Kitáb-i-Ìqàn
Bahá'u'lláh	Kalimát Maknùneh Fàrsi
Bahá'u'lláh	Kalimát Maknùneh 'Arabi
Bahá'u'lláh	Majmù'eh Àlvàh b'ad az Kitáb-i-Àqdas

# Bibliography

- [1] (). Persian digital library by PersDigUMD, [Online]. Available: <https://persdigumd.github.io/PDL/> (visited on 08/21/2019).
- [2] M. Romanov and M. Seydi, *OpenITI: A machine-readable corpus of islamicate texts*, type: dataset, May 20, 2019. DOI: 10.5281/zenodo.3082464. [Online]. Available: <https://zenodo.org/record/3082464#.XWQ6jpNKh24> (visited on 08/26/2019).
- [3] S. Jaf, “A semi-automatic approach to identifying and unifying ambiguously encoded arabic-based characters,” in *2016 International Conference on Asian Language Processing (IALP)*, Nov. 2016, pp. 228–231. DOI: 10.1109/IALP.2016.7875974.
- [4] N. Saiedi, *Gate of the heart: understanding the writings of the Báb*, ser. Bahá’í studies series. Waterloo, Ont.: Wilfrid Laurier University Press, 2008, 423 pp., ISBN: 978-1-55458-035-4.
- [5] Bahá’u’lláh, *The Hidden Words*. Baha’i Publishing, 2002, 191 pp.
- [6] ———, *Tabernacle of Unity*. 2006. [Online]. Available: <https://www.bahai.org/library/authoritative-texts/bahauallah/tabernacle-unity/tabernacle-unity.pdf?e807341d> (visited on 11/04/2019).
- [7] Bahá’u’lláh and Muhyi’-d-Din Sabri Kurdi Sanandaji Kanimishkani, *Majmu’ih-’i Matbu’ih* (“*Collected Tablets of Bahá’u’lláh*”). Cairo: Matba’at as-Sa’adah, 1920.
- [8] Bahá’u’lláh, *Muntakhabati-az Asar-i-Hadrat-i-Bahá’u’lláh*. 163.
- [9] M. Gail, “World order,” vol. 12, p. 48, Summer 1978 1978. [Online]. Available: <https://www.kahlilgibran.com/digital-archive/165-juliet-thompson-remembers-gibran-as-told-to-marzieh-gail-world-order-12-4-1978-pp-29-31/file.html> (visited on 11/07/2019).
- [10] M. Momen, Ed., *Selections from the Writings of E. G. Browne on the Babi and Bahá’i Religions*, Oxford, UK: George Ronald, 1987.
- [11] Lewis, Frank, “Scripture as literature: Sifting through the layers of the text,” *Bahá’i Studies Review*, vol. 7, 1997. [Online]. Available: [https://bahai-library.com/lewis\\_scripture\\_literature](https://bahai-library.com/lewis_scripture_literature) (visited on 10/25/2017).
- [12] W. McCants, “The wronged one: Shí’í narrative structure in bahá’u’lláh’s tablet of visitation for mullá husayn,” p. 12, 2002.
- [13] Lawson, Todd, “Globalization and the hidden words,” in *Bahá’i and Globalization*, Warburg, Margit, Hvithamar, Annika, and Warmind, Morten, Eds., Denmark: Aarhus University Press, 2005, pp. 35–53.
- [14] Justin Rice. (Jan. 3, 2018). What makes hemingway hemingway? Litcharts.com, [Online]. Available: <https://www.litcharts.com/blog/analytics/what-makes-hemingway/> (visited on 11/16/2019).

- [15] A. Kilgarriff, “Comparing corpora,” *International Journal of Corpus Linguistics*, vol. 6, no. 1, pp. 97–133, Dec. 31, 2001, ISSN: 1384-6655, 1569-9811. DOI: 10.1075/ijcl.6.1.05kil. [Online]. Available: <http://www.jbe-platform.com/content/journals/10.1075/ijcl.6.1.05kil> (visited on 08/26/2019).
- [16] F. D. Laramée, “Introduction to stylometry with python,” *Programming Historian*, Apr. 21, 2018. [Online]. Available: <https://programminghistorian.org/en/lessons/introduction-to-stylometry-with-python> (visited on 12/31/2019).
- [17] Genette, Gérard, *Palimpsests: Literature in the Second Degree*, ser. Stages volume 8. Lincoln and London: University of Nebraska Press, 1997.