BRILL

Arabica 58 (2011) 210-299

Arabica

brill.nl/arab

The Chronology of the Qurʾān: A Stylometric Research Program

Behnam Sadeghi[1]

Abstract

I verify a chronology in which seven groups of passages represent consecutive phases. A proposed chronology is verified if *independent* markers of style vary over its phases in a smooth fashion. Four markers of style follow smooth trajectories over the seven phases: The first is average verse length. The second encompasses the 28 most common morphemes in the Qurʾān. The percentages of these morphemes in a text constitute its stylistic profile. The thus-defined stylistic profile is shown to vary in a smooth fashion over "time", *i.e.* over the proposed chronological sequence of phases. Third, a similar thing holds for a profile based on the frequencies of 114 other common morphemes. Fourth, similar results are obtained for a list of 3693 relatively *un*common morphemes. In addition to establishing a relative chronology in seven phases, this essay demonstrates the stylistic unity of many large passages. It also shows that the Qurʾān has one author.

Keywords

Qurʾān, *sūra*s, chronology, stylometry, Islamic origins, *Sīra*, Prophet Muḥammad, Mehdi Bazargan

1. Introduction

> And a Koran We have divided,
> for thee to recite it to mankind
> at intervals, and We have sent it down
> successively.[2]

[1] This work was supported through the research fund of Michael Cook at Princeton University in 2005. The essay was presented in Nov. 2006 in the American Academy of Religion Conference in Washington, DC. It was submitted to Arabica in 2009, but was updated before publication with references to recent literature. I thank Michael Cook for his written comments on this essay, Shuly Wintner for generous assistance with the "tagged Qurʾān" he developed jointly with the late Rafael Talmon, Andrei Radulescu for invaluable assistance with computers, Abdolali Bazargan and Mohammad Hossein Bani Asadi for gifting several volumes of Mehdi Bazargan's *Sayr*, and Asad Ahmed and Patricia Crone for useful discussions.

[2] Kor 17, 106. See also Kor 25, 32. Arthur Arberry, *The Koran Interpreted*, London, Oxford University Press, 1983.

DOI: 10.1163/157005810X529692

The Goal

Using stylometry, I answer three related questions: First, how many authors does the Qurʾān have? Second, is the basic textual unit a small fragment normally no more than a few sentences long, or do many relatively large passages form stylistically coherent units? Third, what is the relative order in which the passages of the Qurʾān were disseminated? This last problem, *viz.* relative chronology, has long been a topic of scholarly controversy, and it is what provides the impetus for this essay. Knowing the relative chronology of the Qurʾān is important if one hopes to interpret it properly and use it to understand the formation of Islam. The answers to the first two questions emerge as corollaries of the analysis conducted for chronology.

The analysis here covers the entire Qurʾān. The last time a publication appeared with a similar scope was in 1976, when Mehdi Bazargan (Mahdī Bāzargān) published the first volume of his landmark *Sayr-i taḥawwul-i Qurʾān*.[3] Bazargan's work has inspired mine and provided the starting point

[3] The publication history of the three parts of Bazargan's book requires clarification. The first volume, comprising about 200 pages, was published in HS [Solar Hejri=modern Iranian calendar] 1355/1976-7, many years after it was written (Mahdī Bāzargān, *Sayr-i taḥawwul-i Qurʾān*, vol. I, Tehran, Qalam, HS 1355/1976-7). This volume contained the main results and a quantitative study of the distribution of themes in the Qurʾān. It also contained a summary in the French language. The second volume, offering a *sūra*-by-*sūra* discussion of the block divisions, was published in HS 1360/1981 (Mahdī Bāzargān, *Sayr-i taḥawwul-i Qurʾān*, vol. II, Tehran, Širkat-i Sahāmī-i Intišār, HS 1360/1981). In HS 1362/1983-4 a "complementary volume" (*Mutammim*) was published with no overlap with vols I and II, offering a qualitative discussion of the evolution of ideas and language in the Qurʾān. In HS 1377/1998-9, a revised edition of the first volume was published, expanded to include a 200-page addition comprising the complementary *Mutammim* published earlier and incorporating corrections to the results in the first edition (Mahdī Bāzargān, *Sayr-i taḥawwul-i Qurʾān*, vol. I, Tehran, Širkat-i Sahāmī-i Intišār, HS 1377/1998-9, 428 p.). In this edition, the citations of page numbers in the French part were not updated. Finally, in HS 1386/2007, a new edition of the book was published in which a group of researchers used computers to make corrections (Mahdī Bāzargān, *Sayr-i taḥawwul-i Qurʾān*, vols I and II in one volume, Tehran, Širkat-i Sahāmī-i Intišār, HS 1386/2007, 617 p.). This edition includes the revised versions of the old volumes I and II in one volume, but does not include the contents of the abovementioned complementary *Mutammim* volume. Unfortunately, volume I in this latest edition is missing some materials included in the previous editions. This includes, for example, the final section, corresponding to p. 157-210 in the 1976-7 edition and p. 181-234 in the 1998-9 edition, which is devoted to plotting the distribution of different themes against time.

In my citations of vol. I, three page numbers are given separated with slashes, referring respectively to the editions of 1355/1976-7, 1377/1998-9, and 1386/2007.

In my citations of vol. II, two page numbers are given separated with slashes, referring respectively to the editions of 1360/1981 and 1386/2007.

For the *Mutammim* volume, I use the abbreviation *Sayr Mutammim*, and the page numbers refer to the 1377/1998-9 edition.

for my analysis. Its focus on style and its quantitative cast are methodological features that carry over into my work.

However, four features distinguish my contribution from the works of Bazargan and other researchers. First, this essay answers the above three questions (*viz.* the number of authors, the basic textual unit, and chronology) without any recourse to the statements of early Muslims about the history of the Qurʾān. It uses neither the reports of individuals about Islamic origins nor the broad historical framework taken for granted in the huge literature that these reports comprise collectively. The reliability of such statements has been the subject of an academic debate. Focusing on the style of the Qurʾān, this article bypasses the sayings of early authorities altogether and therefore is immune to doubts regarding their authenticity and reliability. In fact, the results here constitute an independent test of the broad outline of Islamic beginnings given in the literary sources.[4]

The second contribution involves the distinction between two things: (1) the criterion used for generating a sequence of groups of passages that one conjectures as representing the chronological order, and (2) the criterion one uses to corroborate or verify that sequence. Bazargan and other scholars used style to generate such sequences, but largely failed to use style to corroborate them. Their sequences could claim corroboration from considerations of meaning and external literary evidence—the kind of sources I disregard for the purpose of this article, but they were for the most part uncorroborated by style.[5] By contrast, this essay provides a purely style-based method of verifying

---

My analysis of Bazargan's chronology is based on the original 1976-7 edition and does not take into account the corrections that were introduced later. This is because when I finished my analysis in 2005, I did not yet have access to the other volumes. My study concerns broad and robust patterns, so the kinds of small deviations introduced in later editions are immaterial.

[4] This is not to say that I am neutral in the debate. I recognize the difficulty of evaluating the historical reports bearing on chronology. For example, the reports on the occasions of revelation (*asbāb al-nuzūl*) are often contradictory and speculative. But I do not grant the plausibility of the revisionist claims that (1) the study of the literary sources *cannot* shed light on Islamic origins and that (2) the *broad outline* of Islamic history found in those sources is unreliable. See Behnam Sadeghi, "The Codex of a Companion of the Prophet and the Qurʾān of the Prophet", *Arabica*, 57/4 (2010), p. 343-436. The point is that the method in the present essay circumvents the problems and debates associated with the literary sources.

[5] These statements apply to pre-modern scholars, Nöldeke, and Bazargan. Some premodern authorities distinguished two phases using stylistic elements, including word choice. See Rāmyār, *Tārīḫ-i Qurʾān*, p. 604-9; Muḥammad b. ʿAbd Allāh al-Zarkašī, *al-Burhān fī ʿulūm al-Qurʾān*, ed. Abū l-Faḍl al-Dimyāṭī, Cairo, Dār al-ḥadīṯ, 1427/2006, p. 132-4. For example, some said that the words *kallā* and *yā ayyuhā l-nās* occur in Meccan *sūra*s, while *yā ayyuhā llaḏīna āmanū* occurs in Medinan *sūra*s.

For the Weil-Nöldeke chronology, verse length was a key criterion. See Gustav Weil, "An Introduction to the Quran. III" (translated by Frank Sanders, *et al.*), *The Biblical World*, 5/5

proposed sequences, and it applies this method to corroborate a seven-phase chronology. The approach is rooted in what I call the "Criterion of Concurrent Smoothness". The criterion does not generate chronological sequences; rather, it judges sequences that have been derived by other means. It does not matter how a proposed sequence was obtained: as long as it satisfies the criterion, it is confirmed as genuinely chronological.

Third and related, previous style-based proposed sequences were generated based on the assumption that the style of the Qurʾān changed in one direction without reversals. This is the way certain pre-modern authorities thought about the classical binary Meccan-Medinan division. It is also the approach taken in the four-phase Weil-Nöldeke chronology and the more detailed chronology of Bazargan. Bazargan assumed that verse length tended to increase over time, without reversing course, and he used this principle to rearrange the passages. By contrast, seeking to approach the problem with minimal premises, I make no *a priori* assumptions about how the style of the Qurʾān developed over time. I do not presuppose that style must have progressed irreversibly, nor that it must have changed gradually. I do not even assume that style *must* have evolved in a manner satisfying my criterion for evaluating chronological sequences, namely the Criterion of Concurrent Smoothness. Thus, if a sequence does not satisfy this criterion, that does not mean that it is not the true chronological sequence. But if it does satisfy it, then it is.

Fourth, in the literature there are several quantitative investigations of some aspects of the Qurʾān.[6] These contributions are valuable, but relatively limited

---

(May 1895), p. 343-59; *idem*, "An Introduction to the Quran. IV", *The Biblical World*, 5/6 (June 1895), p. 438-47; Theodor Nöldeke, *Geschichte des Qorāns*, 2nd ed., ed. Friedrich Schwally, *et al.*, Hildesheim, Olms, 1961, p. 58-234. Nöldeke's *Geschichte* has been translated by George Tamer into Arabic as *Tārīḫ al-Qurʾān*, Beirut, Konrad-Adenauer-Stiftung, 2004.

In a recent essay, Nicolai Sinai takes the first phase (early Meccan period) of the Weil-Nöldeke chronology (corresponding approximately to the first three groups in Bazargan's chronology) and breaks it up into four sub-phases: I, II, IIIa, and IIIb. The last sub-phase (IIIb) is distinguished from the previous three by longer verses. In this respect, Sinai's work mirrors that of Bazargan. However, Sinai also considers *sūra* length as a criterion. The four sub-phases are arranged in order of increasing *sūra* length. For the first three sub-phases it is not possible to claim a convergence of the criteria of verse length and *sūra* length, since their verse-length profiles are similar. However, if one combines the first three sub-phases (I, II, IIIa) into one subphase, the outcome is two sub-phases that exhibit a convergence of the two stylistic criteria of verse length and *sūra* length. Sinai also considers the number of sub-sections in a *sūra*, a variable that is probably not independent from *sūra* length. See Nicolai Sinai, "The Qur'an as Process", in *The Qurʾān in Context: Historical and Literary Investigations into the Qurʾānic Milieu*, ed. Angelika Neuwirth, *et al.*, Brill, Leiden, 2010, p. 407-40.

[6] Mehdi Bazargan has done the most substantial work and he is the only one to use quantitative methods to synthesize a chronological sequence. However, he does not use the techniques of multivariate statistics. His book is cited in footnote 3.

in either scope or methodology. On the one hand, studies conducted by scholars like Bazargan who know the historical problems are circumscribed in their use of quantitative techniques. On the other hand, researchers with the best statistical and computer skills lack a historical and literary background, which hampers them in formulating questions that advance Qurʾānic Studies. The present article helps close the gap. It is a systematic analysis of the entire Qurʾān that uses some methods that are commonly used in the field of stylometry but which have not been applied to the study of the Qurʾān before (*e.g.* principal component analysis). The essay also uses some methods that are *not* common in stylometry, such as the cutting-edge technique of weight optimization. It offers a statistical treatment of many stylistic features: common words and morphemes, function words, uncommon words, word length, verse length, and hapax legomena. It needs to be stressed, however, that I have taken

---

Naglaa Thabet uses multivariate techniques, namely hierarchical clustering. She compares the vocabulary of twenty-four long *sūra*s, excluding function words. This reveals two major clusters of *sūra*s which she identifies with the Meccan-Medinan division. See Naglaa Thabet, "Understanding the thematic structure of the Qur'an: an exploratory multivariate approach", *Proceedings of the ACL Student Research Workshop*, 2005, p. 7-12. Hermann Moisl follows up on the work of Thabet. He discusses the limitations that *sūra* length places on meaningful clustering. He clusters 47 *sūra*s on the basis of the frequencies with which they use nine lexical items (*llāh, lā, rabb, qāla, kāna, yawm, nās, yawmaʾiḏin*, and *šarr*). He does not discuss chronology beyond the Meccan-Medinan division. See Hermann Moisl, "Sura Length and Lexical Probability Estimation in Cluster Analysis of the Qur'an", *ACM Transactions on Asian Language Information Processing (TALIP)*, 8/4 (Dec. 2009), article 19.

Nora Schmid has written an essay that has two parts. The first part shows that in Nöldeke's four-phase chronology the *sūra*s in each phase tend to have longer verses than those in the preceding phase. (Bazargan had made a similar observation.) This is to be expected since verse length was one of the criteria used to define the phases in that chronology. More significantly, she shows that in the Meccan period average verse length is somewhat correlated with *sūra* length. Thus, verse length as a criterion seems, at least in the Meccan period, somewhat corroborated by *sūra* length. See also the essay of Nicolai Sinai cited in the previous footnote. Note that the idea that in the Meccan period the length of a *sūra* is related to its chronological rank is implicit in Iʿtimād al-Salṭana's chronological sequence (see below, footnote 24), which lists Meccan *sūra*s in the reverse order of their place in the ʿUṯmānic Qurʾān, hence roughly in order of increasing *sūra* length. The second part of Schmid's essay shows that in the Middle Meccan *sūra*s, a short verse tends to be adjacent to a short verse, and a long verse tends to be adjacent to a long verse, thus refuting the hypothesis that individual verses were put next to each other in a random fashion. Whether this entails coherence at the *sūra* level requires more discussion. See Nora Schmid, "Quantitative Text Analysis and the Qurʾān", *The Qurʾān in Context: Historical and Literary Investigations into the Qurʾānic Milieu*, ed. Angelika Neuwirth, *et al.*, Brill, Leiden, 2010, p. 441-60.

Hans Bauer considered the extent to which the arrangements of the *sūra*s in the ʿUṯmānic codex and the codices of Ibn Masʿūd and Ubayy b. Kaʿb depend on the lengths of the *sūra*s. See Hans Bauer, "Über die Anordnung der Suren und über die geheimnisvollen Buchstaben im Qoran", *Zeitschrift der Deutschen Morgenländischen Gesellschaft*, 75 (1921), p. 1-20.

pains to make the treatment accessible to historians who have not studied statistics and hardly even remember any math from high school.

Beyond what has been accomplished, the prospect of what can be done in future is exciting. The corroboration of a seven-phase chronology may be gratifying, as it turns a sequence that is hypothetical into one that is verified, and as it nearly doubles the number of the phases in the Weil-Nöldeke chronology. However, there is reason to believe that the methods used here, especially the Criterion of Concurrent Smoothness and weight optimization, will enable future stylometric studies to verify more precise chronologies, increasing the number of phases beyond seven. Before us now lies the vista of a new research program in the chronology of the Qurʾān.

Motivating the Approach in This Essay

I call the seven-phase chronology corroborated here the "Modified Bazargan" chronology. Figure 1 shows how this sequence stands in relation to some others. The Weil-Nöldeke chronology divided the Meccan *sūra*s into three phases, yielding a total of four phases. Bazargan offered a more detailed chronology that is portrayed here as twenty-two phases. The Modified Bazargan sequence is obtained by combining some consecutive phases of Bazargan. For example, the passages in the last three phases of Bazargan are combined into a single phase in the Modified Bazargan chronology. This results in a reduction from twenty-two to seven phases as shown. It is necessary to begin with a discussion of Bazargan's work.

Bazargan divides the 114 *sūra*s of the Qurʾān into 194 blocks, preserving some *sūra*s intact as single blocks while dividing others into two or more blocks. He then rearranges these blocks approximately in order of increasing average verse length. This order, he proposes, is the chronological order. His working assumption is that over time the style of the Qurʾān, as represented by verse length, changed gradually—indeed not only gradually but also monotonically, *i.e.* irreversibly in one direction. He stresses that his proposed chronology should not be taken as rigid because it is statistical in nature and because statistical methods sustain firm conclusions about averages of aggregates rather than individual items.

Bazargan corroborates his proposed chronology in a number of ways. He points to broad agreement with Blachère's chronology, which is almost the same as Nöldeke's redaction of Weil's phases.[7] In addition, he examines fifteen instances where historical information suggests a date for a passage. In his

[7] Bāzargān, *Sayr*, vol. I, p. 25 / 43 / 50; and p. 100-13 / 122-35 / 128-45.

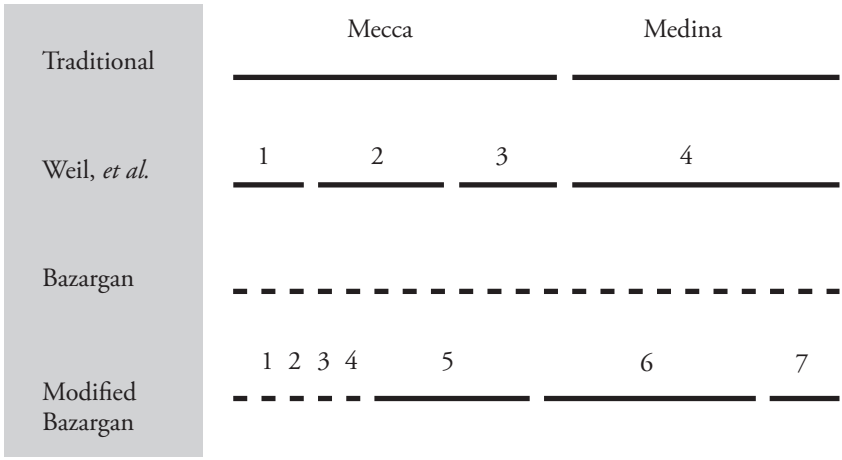

Figure 1. The Modified Bazargan chronology is obtained by combining some consecutive phases of the Bazargan sequence to yield seven phases. For example, phase 7 consists of all the passages in the last three phases of Bazargan (20-22). Corroboration is not claimed for Bazargan's first phase, which contains only 415 words.

chronology, thirteen of these passages line up in the expected sequence.[8] I performed a similar test using nine Medinan passages discussed by Neal Robinson, including three passages not found in Bazargan's list.[9] Bazargan's chronology puts seven of them in the expected order. As another test, Bazargan traces the development of themes over time. He notes, for example, that the verses on wine line up in the expected way, in order of increasing severity.[10] Furthermore, his chronology divides the Qurʾān into two halves whose thematic profiles fit the Meccan and Medinan phases of the Prophet's career. For example, with only a few exceptions, the theme of war crops up in the second (Medinan) half of Bazargan's chronology.[11]

---

[8] *Ibid.*, p. 127-36 / 149-55 / 160-9.

[9] Neal Robinson, *Discovering the Qurʾān: A Contemporary Approach to a Veiled Text*, 2nd ed., Washington, D.C., George Washington University Press, 2003, p. 37-44.

[10] Bāzargān, *Sayr Mutammim*, p. 239-42. He also discusses *ribā*, sexual norms, war and *ǧihād*, the Hypocrites, Abraham, man's creation, spiritual purification and *zakāt*, and Moses and the Children of Israel; see *Sayr Mutammim*, p. 243-409.

[11] The exceptional war-related verses in the Meccan part of Bazargan's chronology are as follows: Kor 2, 244 (Block 113), Kor 22, 58 (Block 124), and Kor 73, 20 (Block 132). These are surely Medinan. See Bāzargān, *Sayr*, vol. I, p. 183 / 207 / missing; and *Sayr Mutammim*, p. 258-9. I have not included references to *ǧihād* among the exceptions, since they need not be

The agreement with Blachère's chronology is to be expected to the extent that the latter uses verse length as a criterion. But that chronology relies also on other criteria and sources, such as early traditions and considerations of meaning. This suggests that there is more to Bazargan's reasoning than circular justification. Furthermore, the backing provided by agreement with the historical information is very impressive and completely devoid of circularity.

While Bazargan orders the blocks using a stylistic criterion, in order to support the validity of his ordering, he reaches beyond style, adducing considerations of meaning and historical evidence. Ultimately, semantic and historical evidence should play an indispensable role in the evaluation of any proposed chronology. However, it is useful to ask how far one can go in corroborating a chronology using *only* style. The advantage of initially limiting oneself to style is that subsequently it may enable independent corroboration in a more meaningful fashion. That is, to the extent that style-based indications agree with other forms of evidence, one can speak of joint corroboration by genuinely independent strands of evidence. Yet if from the outset one mixes different kinds of analysis, then in the end it may be less straightforward to tell independent corroboration apart from the duplication of the same evidence in different guises. There is, therefore, merit in *initially* preventing semantic and historical considerations from shaping one's analysis of chronology.

The Approach in This Essay

The present task, then, is to see to what extent one can corroborate Bazargan's chronology through purely stylistic considerations. The first step is to ask how much corroboration his chronology enjoys at present if its semantic and historical justifications were stripped away. The answer would have to be: relatively little. Although it may be questioned, his postulate that the style of the Qurʾān changed gradually is not implausible. But he does not stop at that, as he also assumes that it evolved only in one direction. But why should average verse length have increased monotonically? Why, for example, could it not have increased for some time, slowly leveled off, and then decreased gradually for a while more, before leveling off again or resuming an upward trend? Indeed, one can take the Qurʾān, or any other corpus of texts, and rearrange it in many different ways that all make a particular marker of style, such as sentence length, change in a relatively continuous manner. Yet, most such orderings will not reflect chronology.

---

understood in the military sense. For chronological graphs of other themes (eschatology, past prophets, the People of the Book, etc.), see *Sayr*, vol. I, p. 176-96 / 201-20 / missing.

Clearly, then, a single marker of style cannot strongly corroborate any particular chronology. But what if one took into account additional markers of style: the relative frequencies of the most common words, the distribution of word lengths, and so on? In general, given a corpus of texts, it would appear impossible to rearrange it to make many independent markers of style vary gradually, as opposed to in a jagged fashion. This is because as soon as one rearranges a text to make one marker of style vary smoothly, that will affect the behaviors of all the other markers, interfering with their smoothness. Finding an arrangement that makes all variations smooth would be quite extraordinary. Such a pattern could not be due to chance, thus requiring an explanation. Chronological development would usually be the only plausible explanation. That this would be *a* natural explanation must be obvious enough: if style did indeed change gradually, continuous change could very well characterize more than one marker of style (though it would not have to), and the chronological sequence would thus make multiple markers of style vary smoothly. That this is usually the only plausible explanation reflects our inability to adequately explain the pattern in other ways.[12] Concurrent smoothness, when observed, cannot be a coincidence; thus, potential critics of my approach have the burden of explaining it without resort to chronology.

Methodology: The Criterion of Concurrent Smoothness

In sum, the principle underlying my study is that if different, independent markers of style vary in a relatively continuous fashion over a particular ordering, then that sequence reflects the chronological order. The point is that while it is easy to find many orderings of a corpus over which one particular marker of style varies smoothly, it is highly unlikely that an ordering will yield smooth variation simultaneously for different, independent markers of style ("concurrent smoothness"). This is so because if one reorders the corpus to make one

[12] There is a caveat and a potential objection. The caveat: There is a way to artificially construct a sequence that makes different markers smooth without the sequence reflecting chronology. Mixing two different, discrete styles in gradually changing proportions can accomplish that. The Conclusion explains this possibility and discusses whether it characterizes Bazargan's sequence. The potential objection: This essay shows that average verse length correlates with morpheme frequencies. It may be objected that this correlation reflects facts of linguistics: certain grammatical structures can be used more easily in shorter sentences. In response, I note two things. First, the person who makes this objection has the burden of demonstrating the claim. Second, it would not be implausible for such a phenomenon to affect a few morphemes; but the results of this essay do not change if several morphemes are deleted from the lists of morphemes used. The essay uses three lists including respectively 28, 114, and 3693 morphemes; and only those conclusions are accepted that are confirmed by all three lists or at least two of them.

marker of style vary smoothly, usually this reordering will disturb the smoothness of other markers. In the unlikely event that one does find a reordering that achieves concurrent smoothness, usually the only plausible explanation is that it reflects the gradual variation of different markers of style over time.[13] *I thus examine whether markers of style other than verse length also vary in a smooth fashion over Bazargan's sequence. If they do, and to the extent that they do, that confirms this ordering as a reflection of chronology*.

Note that my approach does not take it as a *premise* that as the Qurʾān was revealed, its style changed gradually. If in reality there were major breaks in the style of the Qurʾān over time, then that would simply mean that one would not be able to find an ordering that achieves concurrent smoothness. Where I cannot find such a sequence, I simply conclude nothing.[14] I do not begin with a working assumption of gradual change, even though such a hypothesis would not be implausible.

The reader may have noted a certain asymmetry in my approach: concurrent smoothness confirms a chronology, but the converse does not hold. If an arrangement does not achieve concurrent smoothness, while that fact is certainly suggestive, I would be reluctant to conclude that the order is not chronological, preferring instead to suspend judgment. Gerard Ledger takes a similarly asymmetric approach to the authenticity of disputed Platonic texts. If a disputed text is stylistically identical or close to undisputed works, that is evidence of authenticity; but if a disputed text is stylistically atypical, "that in itself does not constitute proof that it was written by another, unless it can be shown that the author in question never departed from his standard established style".[15] Ledger's cautiously asymmetric approach arises from his unwillingness to make gratuitous assumptions about Plato's manner of writing. Similarly, the asymmetry in my approach is meant to minimize preconceptions about how the Qurʾān's style changed over time. To take lack of smoothness as a sign that a sequence is not chronological would be tantamount

[13] Of course, if one sequence achieves such smoothness, others that are broadly similar to it will do so as well. So, the claim of uniqueness of sequences yielding concurrent smoothness must be understood as valid only within the range of broad similarity.

[14] There is one type of break, however, that if it did occur in reality, would lead to error in my approach. I have in mind a discontinuous reversion to the style of a specific moment in the past, with respect to all markers of style (vocabulary choice, verse length, etc.). I discuss this possibility below in Section 9 ("The Conclusion").

[15] Gerard Ledger, *Re-counting Plato: A Computer Analysis of Plato's Style*, Oxford, Clarendon Press, 1989, p. 168-9. However, one may corroborate that a disputed work is ascribed incorrectly to an author if its style is shown to be identical to that of a known alternative author who is known as a plausible candidate for its authorship; but, even here, considerable caution is required.

to positing gratuitously that the Qur'ān's style *must* have varied smoothly over time.

In sum, concurrent smoothness confirms a proposed sequence as chronological, but lack of concurrent smoothness does not discredit a proposed sequence. This is so because there is no reason *a priori* for assuming that a chronological sequence will yield concurrent smoothness. This does not mean, however, that proposed sequences are unfalsifiable. A proposed sequence is disconfirmed if a different sequence with which it is incompatible is confirmed by concurrent smoothness.

This asymmetry may occasion discomfort among those familiar with Karl Popper's philosophy. Popper believed that a theory may be corroborated only to the extent that it passes a test that could have potentially falsified it. Thus, if a type of evidence cannot potentially falsify a theory, then it cannot corroborate it either.[16] It would appear, then, that if the absence of concurrent smoothness does not refute a proposed sequence, then its presence cannot confirm a sequence either. In response, I note that Popper's observation applies strictly to theories of the form "All A are B", *e.g.* "All swans are white". The situation is often different. Take the following mundane example: if two broken, jagged shards of window glass fit each other perfectly, that is strong evidence for their being from the same broken window. But if they do not fit, that is not strong evidence against their being from the same window.

Stylometry Demystified

To quantify style, I use techniques from the field of stylometry. Bazargan's book itself is a fine example of a stylometric study, although he knew little about the discipline and did not use multivariate techniques. Stylometry has conceptual similarities to the traditional methods of stylistic analysis.[17] The primary difference lies in its use of quantitative methods, which can provide an additional measure of objectivity and help detect statistically significant patterns that might otherwise be difficult or impossible to discern. The use of computers nowadays has become the second distinguishing characteristic of stylometry, as they have made quantitative approaches immensely more practical. Whereas Bazargan did his counting by hand, thanks to computers I do not have to.

---

[16] Karl Popper, *Realism and the Aim of Science*, London, Routledge, 1983, p. 235.

[17] For a methodological comparison of stylometric and traditional techniques, see Section 8 below ("Multivariate Analysis (List C): The Generalized Smoking Gun Technique").

There is a wide range of stylometric approaches, arising in part from the diversity of features used as markers of style. Such features may include punctuation marks, consonants and vowels, character strings, syntactical features, parts of speech, rhythm, hapax legomena (once-occurring words), sentence length, word length, and so on. Probably, the most popular approach is to use the relative frequencies of common function words such as "the", "a", "an", "or", "upon", and "it". My study, too, considers lists of common morphemes.

The most common application of stylometry is establishing the authorship of disputed texts. To do so, typically one compares the style of a disputed text with the styles of known authors. Less commonly, scholars have used stylometry to study issues of chronology.[18] The problem in stylometric literature

[18] See *e.g.* B. Brainerd, "The Chronology of Shakespeare's Plays: A Statistical Study", *Computers and the Humanities*, 14 (1980), p. 221-30; Leonard Branwood, *The Chronology of Plato's Dialogues*, Cambridge, Cambridge University Press, 1990; Fazli Can and Jon Patton, "Change of Writing Style with Time", *Computers and the Humanities*, 38/1 (2004), p. 61-82; A. Devine and L. Stephens, "A New Aspect of the Evolution of the Trimeter in Euripedes", *Transactions of the American Philological Associatio*n, 111 (1981), p. 45-64; W.E.Y. Elliott and R.J. Valenza, "Can the Oxford Candidacy be Saved? A Response to W. Ron Hess: 'Shakespeare's dates: the Effect on Stylistic Analysis'", *The Oxfordian*, 3 (2000), p. 71-97; J. Fitch, "Sense-Pauses and Relative Dating in Seneca, Sophocles, and Shakespeare", *American Journal of Philology*, 102 (1981), p. 289-307; Richard Forsyth, "Stylochronometry with Substrings, or: A Poet Young and Old", *Literary and Linguistic Computing*, 14/4 (1999), p. 467-78; Bernard Frischer, *Shifting Paradigms: New Approaches to Horace's Ars Poetica*, Atlanta, Georgia, Scholars Press, 1991; MacD. Jackson, "Pause Patterns in Shakespeare's Verse: Canon and Chronology", *Literary and Lingustic Computing*, 17/1 (2002), p. 37-46; Patrick Juola, "Becoming Jack London", *Proceedings of Qualico-2003*, Athens, Georgia, 2003; Anthony Kenny, *The Computation of Style*, Oxford, Pergamon Press, 1982; Gerard Ledger, *Re-counting Plato: A Computer Analysis of Plato's Style*, Oxford, Clarendon Press, 1989; S. Michaelson and A.Q. Morton, "Things Ain't What They Used to Be: A Study of Chronological Change in a Greek Writer", in *The Computer in Literary and Linguistic Studies, Proceedings of the Third International Symposium*, eds Alan Jones and R.F. Churchhouse, Cardiff, University of Wales Press, 1976, p. 79-84; Charles Muller, *Étude de statistique lexicale: Le vocabulaire de théâtre de Pierre Corneille*, Paris, Larousse, 1967; Debra Nails, *Agora, Academy, and the Conduct of Philosophy*, Dordrecht, Kluwer, 1995; Behnam Sadeghi, "The Authenticity of Two 2nd/8th-Century Ḥanafī Legal Texts: The *Kitāb al-Āṯār* and *al-Muwaṭṭa'* of Muḥammad b. al-Ḥasan al-Shaybānī", *Islamic Law and Society*, 17/3 (Nov. 2010), p. 291-319; J.A. Smith and C. Kelly, "Stylistic Constancy and Change across Corpora: Using Measures of Lexical Richness to Date Works", *Computers and the Humanities*, 36/4 (2002), p. 411-30; Constantina Stamou, *Dating Victorians: An Experimental Approach to Stylochronometry*, Saarbrücken, Verlag Dr. Müller, 2009; D.K. Simonton, "Popularity, Content, and Context in 37 Shakespeare Plays", *Poetics*, 15 (1986), p. 493-510; *idem*, "Imagery, Style, and Content in 37 Shakespeare Plays", *Empirical Studies of the Arts*, 15 (1997), p. 15-20; Tomoji Tabata, "Investigating Stylistic Variation in Dickens through Correspondence Analysis of Word-Class Distribution", in *English Corpus Linguistics in Japan,* ed. Toshio Saito, Junsaku Nakamura and Shunji Yamazaki, Amsterdam, Rodopi, 2002, p. 165-82; J.T. Temple, "A Multivariate Synthesis of Published Platonic Stylometric Data", *Literary and Linguistic Computing*, 11/2 (1996), p. 67-75; David Wishart and

most similar to the Qur'ānic case is perhaps that of the chronology of Plato's dialogues.[19] According to Gerard Ledger, a common shortcoming of such studies of Plato is that they assume that a marker of style must have changed over time in one direction.[20] This is just the kind of questionable premise that my approach, as described above, is designed to circumvent.

The basic idea in stylometry is to use a statistic, such as the frequency of occurrence of a certain feature, as a marker of style. A limited sample of writing is used to gain an estimate of an abstract, hypothetical quantity, namely the frequency with which an author would use a certain feature if he wrote an infinite amount of text. For example, within the 694 characters in this paragraph, the letter "i" occurs forty-nine times: a *relative frequency count* of 49/694 = 7.1 %. Moreover, to get better results, instead of just the letter "i", one may consider many features simultaneously, such as all letters of the alphabet. Doing so would characterize my style by a *vector*, *i.e.* list of relative frequency counts. When one represents texts with vectors instead of single numbers, the data are called *multivariate*.

Observed relative frequency counts may be thought of as approximations to an abstraction that may be called an author's "true style". For example, the observed 7.1 % frequency count in my last paragraph may be thought of as an approximation to my "true style" of, say, 8 to 9 %. The more precise the "true style" defined by the chosen signifier(s) of style, the more effective the method is likely to be. Thus, some markers of style will be better discriminators than others. Furthermore, larger sample sizes are helpful. To get a better estimate of my usage of *i*'s, one should use a larger sample than the mere 141 words in the last paragraph. The greater the sample size is, the smaller the *sampling error*. The sampling error is a product of the uncertainty introduced by the limited size of the sample.[21]

The most immediate and fundamental task of stylometry is to determine how similar two or more texts are in terms of style. Similarity is always relative. One may say, "A is stylistically similar to B"; but one really means, "text A is *more* similar to text B *than* to text C". Furthermore, similarity depends entirely

---

Stephen Leach, "A Multivariate Analysis of Platonic Prose Rhythm", *Computer Studies in the Humanities and Verbal Behavior*, 3 (1970), p. 90-9.

[19] Ledger, *Re-counting Plato*; Nails, *Agora*, p. 97-114; Branwood, *Chronology*; and especially Wishart and Leach, "A Multivariate Analysis of Platonic Prose Rhythm". For the full citations, see the previous footnote.

[20] Ledger, *Re-counting Plato*, p. 175; cf. Nails, *Agora*, p. 113-4.

[21] For an accessible introduction to the basic concepts of (univariate) statistics, see Kenny, *The Computation of Style,* cited above. I also benefited from working through Larry Christensen and Charles Stoup, *Introduction to Statistics for the Social and Behavioral Sciences*, Belmont, California, Brooks/Cole, 1986. Following this essay does not require any background in statistics.

on the marker of style chosen. Suppose the marker of style is the prevalence of the letter *i*. By this marker, if 6 % of the characters in text A are *i*'s, 7 % of characters in B are *i*'s, and 18 % of characters in C are *i*'s, then one may say that A and B are more similar to each other than either is to C. But if one changed the criterion, say to the frequency of the letter *t*, then one might obtain different results.

A keystone of my study, and a very common technique in stylometry, is the graphical representation of similarity. Stylistic dissimilarity may be represented by the spatial metaphor of distance. "Stylistically different" becomes "far", and "stylistically similar" becomes "near". Texts are represented by points on a page. The closer two points are on the page, the more similar their stylistic profiles are. Obtaining distances between texts is straightforward.

The following example illustrates the method, and grasping it is essential for understanding this essay. Suppose that the chosen marker of style is the frequencies of five common morphemes in the Qurʾān: *wa* ("and"), *u* (nominative case ending), *an* (accusative, indefinite case ending), *fa* (punctuation), and *llāh* ("God"). Suppose one wishes to compare the styles of three chapters of the Qurʾān: *sūra* 13 "Thunder", *sūra* 23 "Believers", and *sūra* 36 "Yā-Sīn". In Figure 2, each *sūra* is represented by a row of five numbers representing the relative frequencies of the five morphemes. The heights of the columns represent the morpheme frequencies: the higher a morpheme frequency is, the taller the column representing it. For example, the tallest column in Figure 2 indicates that 7.2 % of the morphemes in *sūra* 13 consist of the nominative case ending *u*. Each row of numbers forms the stylistic profile of a *sūra*. The back row, for example, represents Thunder. Simply glancing at these rows in order to compare the *sūra* profiles, it is evident that the *sūra* Thunder stands apart from the others, and that the *sūra*s Believers and Yā-Sīn are stylistically closer to each other than either is to Thunder. The question is how to translate these relationships into distances.

To obtain the distance between Thunder and Believers, one takes into account each of the morphemes. Let us begin with the leftmost morpheme, *wa*. In Thunder, 6.6 % of the morphemes consist of *wa*. For Believers, the number is 5.3 %. The difference in the heights of the *wa* columns of the two *sūra*s is 6.6 – 5.3 = 1.3. For the *u* morpheme, the difference is 7.2 – 4.5 = 2.7. For *an*, it is 2.4 – 1.6 = 0.8. For *fa*, it is 2.7 – 1.1 = 1.6. Finally, the difference between the percentages of *llāh* in these two *sūra*s is 1.9 – 0.7 = 1.2. An overall measure of the difference between Thunder and Believers is the sum of these individual differences, *i.e.* 1.3 + 2.7 + 0.8 + 1.6 + 1.2 = 7.6. This number is the *distance* between Thunder and Believers. In general, the distance between any pair of texts is calculated in a similar way: for each morpheme, one takes the

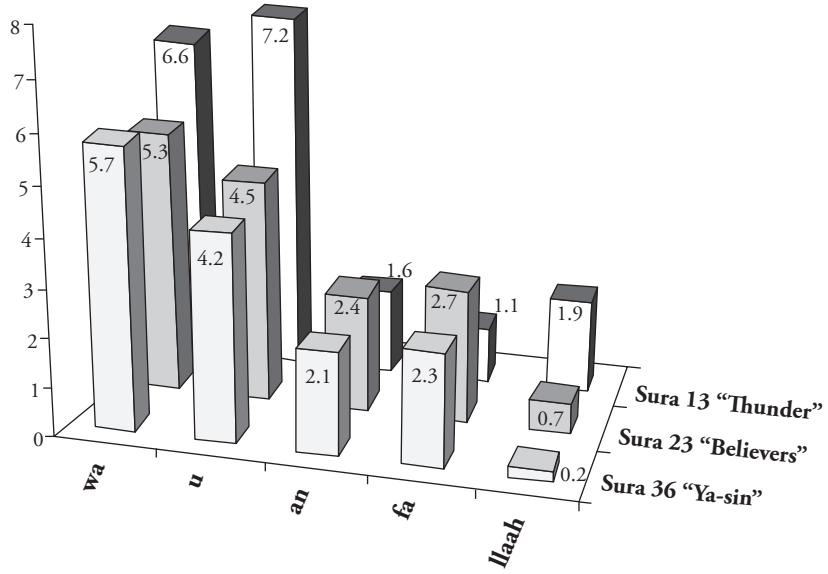

Figure 2. The relative frequency counts of five common morphemes (*wa, u, an, fa*, or *llāh*) in three *sūras, i.e.* their percentages in each *sūra*. The distance between two *sūra*s is found by subtracting their morpheme frequencies and adding up the differences.

difference between its percentages in the two texts, and one adds up the differences for all the different morphemes.

It was just shown that the distance between Thunder and Believers is 7.6. It can be shown, in a similar fashion, that the distance between Thunder and Yā-Sīn is 7.3, and that the distance between Believers and Yā-Sīn is 1.8. These relationships are represented graphically in the following figure. As expected, Thunder stands far apart from the other two *sūra*s.

This diagram is a concise epitome of a great deal of information. The present essay examines up to twenty-two texts at a time based on lists of dozens or even thousands of morphemes; yet the method used to analyze those cases is the same as that in the above example. The graphical method is a convenient and intuitive way of dealing with data of such complexity. The only hitch is that moving from a list of distances to a diagram is not straightforward, as in general it is not possible to reproduce the distances exactly in a two-dimensional picture. Approximations are therefore made using standard methods such as "principal component analysis" to produce diagrams that capture the main trends in the data while filtering out some of the noise.

The above diagram can also be used to illustrate the idea of smoothness. Suppose one wishes to arrange the three *sūra*s in a chronological or reverse-chronological sequence. The three choices are listed here in order of decreasing smoothness

(1) Believers → Yā-Sīn → Thunder
(2) Yā-Sīn → Believers → Thunder
(3) Believers → Thunder → Yā-Sīn

The smoothest sequence is the one in which style progresses most gradually, which is to say that texts that are near in time are stylistically similar. Translated into the language of distances, smoothness means that texts that are close to each other in the chronological sequence are located near each other in terms of stylistic distance. If the texts be thought of as cities and the progression from one text to another as a traveling salesman's itinerary, then the smoothest trajectory is the one that minimizes the total distance traveled by the salesman if he were to stop at every city exactly once. The smoothest path is thus the shortest. In this case, the first sequence above is evidently the smoothest one, the second sequence is slightly less smooth, while the third one is decidedly not smooth. If the first sequence yielded stylistic smoothness for several independent markers of style, that would amount to concurrent smoothness.[22]

That takes care of the main ideas behind the method. Now the reader should be able to interpret most of the diagrams. However, the potential remains for a family of misconceptions. The best way to address them is to deal with an example of the kind of objections such misunderstandings generate: It may be asked, for example, why the different uses of a morpheme are not distinguished. The morpheme *wa* sometimes means "and" and sometimes means

[22] A mathematically precise definition of smoothness is given in the Appendix.

"by" (as in an oath). Surely, the objection goes, it does not make sense to treat the two as the same. This criticism concerns the choices made about what to count. One can readily come up with a long list of other objections in this vein, since every choice involves an element of arbitrariness that may occasion the question, "why not do it in this other way?"

The objection is based on three misunderstandings. The first mistaken assumption is that the plausibility of these various initial decisions is what validates the outcome. Actually, it is the reverse. The validation comes at the end when one confirms that there is concurrent smoothness. As an analogy imagine somebody who picks two jagged shards of glass from a heap of broken windows, shows that they fit perfectly, and claims that they are from the same window. What validates his claim is the improbable fit, not how he went about finding the matching fragments. In fact, had he picked them in a random fashion, his argument would be no less valid.

The second mistaken assumption is that we need to understand how the results follow from the initial choices made. The idea again is best illustrated by an example. Suppose we enter a stadium and see that most people are wearing red. We know immediately that there has to be a reason for it other than chance. We may thus put forward an explanation, *e.g.* that the crowd is showing support for a soccer team, or that this is a political rally organized by communists. It would not undermine such an explanation to point out that Maryam who is in the stadium is in fact wearing red for an entirely different reason. What supports the explanation is the statistical significance of the overall pattern. Individual cases need not fit the pattern. By the same token, it is not necessary for us to understand the role played by individual morphemes such as *wa*. A person making this objection might be surprised to know that some important stylometric studies rely on the frequencies of the letters of the alphabet. Thus, the "s" in "sin" is treated as the equal of the "s" in "soup", and the word "funeral" can be replaced with "real fun" without consequence.

The third mistake is to assume that if the initial decisions (such as the choice of morphemes) are poor, one will end up with a wrong result. In reality, what the initial decisions determine is not the accuracy of the outcome, but its precision. Bad choices will lead to fewer phases or none satisfying concurrent smoothness. Thus, either the conclusions will be less precise or one will arrive at no conclusion. But a less precise result is not necessarily wrong. A two-phase chronology is not "less correct" than a three-phase one. It is like using a magnifying glass instead of a microscope: it may capture much less detail, but what it reveals is not wrong.

The Organization of the Essay

This outline may help readers navigate the rest of this essay by highlighting the most important points. Section 2 ("Bazargan's Chronology") defines twenty-two selections or "groups" of Qurʾānic passages, numbered 1 through 22, which according to Bazargan's chronology were disseminated by the Prophet in the order enumerated. The ultimate goal is to see how various markers of style behave over this sequence and whether they exhibit a smooth trajectory.

Section 3 ("Univariate Assessments of Smoothness") examines several univariate markers of style for smoothness over Bazargan's sequence of groups, *e.g.* average word length and the frequencies of hapax legomena. The reader must examine Figure 4, which shows the variation of average verse length over the sequence of the twenty-two groups defined in Section 2. Noteworthy, too, is Figure 8, which shows a strikingly smooth initial trajectory for the percentages of hapax legomena.

The central task of this article is to see how *multivariate* markers of style vary over the twenty-two phases. Before that task is accomplished, however, there are the two intervening Sections 4 and 5. Historians might find some of the details in these sections difficult, but the larger ideas should be clear.

Section 4 offers a "Non-technical Introduction to Multivariate Methods". The discussion above, under "Stylometry Demystified", is even less technical, and unlike Section 4 it is necessary that the reader come to understand it thoroughly.

Section 5 ("Morphemes: Weighting and Weight Optimization") hones the multivariate techniques. It tests whether stylistic profiles based on morpheme frequencies succeed in showing the stylistic similarity of texts that we may expect in advance to belong together. The texts that are chosen for this purpose are halves of large passages. Morpheme frequencies indeed assign the halves to each other, which shows not only the reliability of the method, but also the stylistic coherence and unity of large passages. The section also explains that one may choose to not accord all morphemes equal weight when forming stylistic profiles of passages and calculating the distances between them. It introduces *weight optimization*, a method for finding the best weights automatically. Being a little involved, Section 5 may be skimmed by casual readers, except that Table 5 and Table 6 require attention, as do the concluding words of the section.

The three sections that follow constitute the heart of this essay: Sections 6, 7, and 8 use three different, independent multivariate markers of style based on morpheme frequencies. Each section examines whether the stylistic profile based on one particular list of morphemes varies in a smooth manner over

the sequence of twenty-two groups. Section 6 does this for the top twenty-eight most frequent morphemes in the Qurʾān. Section 7 considers 114 other common morphemes. Section 8 uses a list of 3693 relatively *un*common morphemes.

The results from the three independent multivariate markers ought to be compared in order to assess the degree of concurrent smoothness. Section 9 ("The Conclusion") accomplishes that. It finds that if some of Bazargan's groups are combined in accordance with Figure 1 above, then all the three different markers of style will exhibit smooth trajectories. Specifically, the following sequence of seven clusters yields concurrent smoothness and hence represents the true chronological order: {Group 2}, {Group 3}, {Group 4}, {Group 5}, {Groups 6-11}, {Groups 12-19}, {Groups 20-22}. The claim is not that the passages in one cluster all came after those in the preceding clusters, but that only *on average* they did so. In addition, the chronology of the passages *within* a cluster is indeterminate. For example, I have not confirmed or refuted that passages in Group 8 on average came after those in Group 7, since the two groups belong to the same cluster. The upshot is that the first half of Bazargan's chronology is broadly confirmed. Its second half, consisting of Groups 12-22, which happen to correspond to Medina in the traditional reckoning, remains largely unconfirmed, although it is at least clear that it comes after the first half.

2. Bazargan's Chronology

Mehdi Bazargan (d. HS 1373/1995) was one of the pillars of Islamist and democratic thought in Iran.[23] His interest in chronology arose during Qurʾān

[23] Mehdi Bazargan was born in HS 1286/1907 in Tehran. He was a professor of thermodynamics and Dean of the Engineering Faculty in Tehran University, where, incidentally, he was the first to establish group prayers on campus. In HS 1329/1950-1, appointed by Muḥammad Muṣaddiq as the first Iranian chief executive of the National Iranian Oil Company, he helped end British control with minimal disruption to the industry. As head of the Water Organization in HS 1332/1953-4, he brought running water to Tehran for the first time. He spent 1963 to 1967-8 (HS 1341 to 1346) in prison for criticizing the Shah. Bazargan became the first prime minister of the Islamic Republic in HS 1357/1979, but resigned after nine months, went into opposition and served a term in the Parliament. Bazargan spent the last years of his life working in private industry. He also continued his religious scholarship and political activism until his death on January 20, 1995 (HS 10/30/1373). The Qurʾān lay at the center of his spiritual, political, and intellectual life. It not only provided the framework and language of much of Bazargan's thought, but also formed the subject of a number of his important works, including a work of exegesis (*tafsīr*) that he arranged according to the chronological order of the Qurʾān. The above biographical information is based mostly on Mahdī Bāzargān, *Ḫātirāt-i Bāzargān:*

study sessions held with his colleagues, mostly from his Freedom Movement, in prison in Burāzğān in the Ramaḍān of HS 1344/1965-6, and it culminated in the completion of the book in Qaṣr Prison. He began his study of verse length upon encountering a chronological list offered in a Qurʾān printed by Iʿtimād al-Salṭana (d. 1330/1912). This list gives a year-by-year break down of the *sūra*s.[24] Bazargan used this list to construct a graph of the mean length of verses versus time, obtaining a rather jagged graph with an overall increasing trend. Having also made graphs of themes vs. time, he noted that some of the breaks in the trends in the two types of graph went together. He then rearranged the *sūra*s to make mean verse length increase monotonically. This removed the sharpest breaks from the plots of the themes and improved agreement with the known provenance of the *sūra*s as Meccan or Medinan. In other words, he had encountered the correlation between style, content, and historical information. Encouraged, he further refined his chronology by breaking up the *sūra*s into blocks and reordering these blocks instead of whole *sūra*s.[25] Eventually, some secondary sources were sent to him in prison, including a copy of Blachère's *Introduction* which allowed him to make comparisons

---

Šaṣt sāl ḫidmat wa muqāwamat, Tehran, Muʾassasa-i Ḫadamāt-i Farhangi-i Rasā, HS 1375/1996. I also used the biography on http://www.bazargan.com, a site maintained by Abdolali Bazargan. Citations to Bazargan's work on chronology are given above in footnote 3.

[24] According to Bazargan, the historian and litterateur Muḥammad Ḥasan Iʿtimād al-Salṭana does not give a source for his chronology. I think the chronology was probably his own. Comparing his chronological sequence with the thirteen lists of Muslim and European origin provided by Mehdi Abedi, one finds that it is unique (Michael Fischer and Mehdi Abedi, *Debating Muslims: Cultural Dialogues in Postmodernity and Tradition*, Madison, University of Wisconsin Press, 1990, p. 445-7). Bazargan points out that the subject index (*kašf al-maṭālib*) that Iʿtimād al-Salṭana provided was, in the words of the latter, composed on the basis of an index created by an European scholar, "one of the *ʿulamā* of the *Maġrib*". After quoting this, Bazargan wonders whether the same European scholar, or maybe Iʿtimād al-Salaṭana himself, was not the source of the chronology (Bāzargān, *Sayr*, vol. I, p. 21-3 / 39-41 / missing). In fact, the European author of the subject index was not the source of the chronology. Muḥammad Nūrī and Maḥmūd Rāmyār both identify the source of the subject index as Jules La Beaume, *Le Koran analysé d'après la traduction de M. Kazimirski et les observations de plusieurs autres savants orientalistes*, Paris, Maisonneuve & Cie ("Bibliothèque Orientale", 4), 1878 (see Muḥammad Nūrī, "*Tafṣīl āyāt al-Qurʾān al-ḥakīm*", *Faṣlnāma-i kitābhā-i Islāmī*, no. 7, available online at http://www.i-b-q.com/far/07/article/09.htm; Maḥmūd Rāmyār, *Tārīḫ-i Qurʾān*, 2nd ed., Tehran, Amīr Kabīr, HS 1362/1983, p. 667). La Beaume, however, does not offer any chronological list. Iʿtimād al-Salṭana's list is as follows (note that its Meccan part is simply the known Meccan *sūra*s listed in the reverse of the official order): Mecca: 1, 114-99, 96-67, 56-50, 46-34, 32-25, 23, 21-10, 7-6; Medina: 97, 64, 63, 62, 98, 61, 57, 47, 58, 59, 60, 49, 22, 4, 24, 65, 66, 33, 48, 8, 9, 2, 3, 5.

[25] See Bāzargān, *Sayr*, vol. I, p. 16-25 / 34-43 / 41-52.

with Blachère's version of Nöldeke's chronology.[26,27] Bazargan's book has been received well.[28] As arguably the most impressive work on the subject, it deserves the praise it has received; but a reevaluation is long overdue.

As mentioned above, Bazargan recognizes that a *sūra* may contain material from different periods. He thus divides *sūra*s into blocks and rearranges these blocks rather than the *sūra*s. In this process, Bazargan takes into account the distribution of verse lengths in different blocks by considering their "characteristic curves". Figure 3 depicts the characteristic curves for six different *sūra*s

---

[26] Régis Blachère, *Introduction au Coran*, 2e éd. partiellement refondue, Paris, Besson & Chatemerle, 1959.

[27] My own interest in this research was sparked, first, by Bazargan's work, and, second, by noting in my study of al-Šaybānī's *Muwaṭṭaʾ* that word frequency could correlate with time. See Sadeghi, "The Authenticity of Two 2nd/8th-Century Ḥanafī Legal Texts", cited above in footnote 18. Subsequently, an inquiry about word frequency in stylistics led me to stylometry and its techniques.

[28] See *e.g.* Rāmyār, *Tārīḫ-i Qurʾān*, p. 665. Among Islamicists impressed with the *Sayr*, the most famous would be Āyatullāh Murtaḍā Muṭahharī, who called the book "highly valuable" (Muṭahharī, *Muqaddama-ī bar ǧahānbīnī-i Islāmī*, Qumm, Ṣabā, n.d., p. 194; I owe this reference to Hossein Modarressi). Incidentally, it was on Muṭahharī's recommendation that Āyatullāh Khomeini appointed Bazargan prime minister. Among religious public intellectuals, the most famous admirer of the *Sayr* was the influential and charismatic ʿAlī Šarīʿatī, although he liked the book perhaps for the wrong reason. An excerpt from his overly sanguine letter to Abdolali Bazargan (ʿAbd al-ʿAlī) dating from 1968-9 appears in the preface of the expanded edition of the *Sayr*. For some criticisms of the *Sayr*, see ʿAlī Riḍā Ṣadr al-Dīnī, "*Naẓarī bih Sayr-i taḥawwul-i Qurʾān*", *Kayhān-i Farhangī*, 3/12 (Isfand HS 1365/1986), p. 14-8. The skeptical Ṣadr al-Dīnī writes, "This book has gained a foothold among the *ʿulamā* and scholars, and its deductions have been met with acceptance. As of late, some of its contents are even being printed in the appendices of some Qurʾāns". For another criticism, see the introduction written by Aḥmad Mahdawī Dāmġānī for Muḥammad Riḍā Ǧalālī Nāʾīnī, *Tārīḫ-i ǧamʿ-i Qurʾān-i karīm* (with an introduction by Aḥmad Mahdawī Dāmġānī), Tehran, Našr-i Nuqra, HS 1365/1986, p. xv.

Western scholars who specialize in the Qurʾān do not discuss Bazargan's work, not even mentioning it in a footnote. This is so despite the fact that Bazargan wrote a summary of his work in French and presented his work in Germany. It is not uncommon for Western scholars to overlook Islamic secondary scholarship on Islamic religion. Aside from the issue of language, this is partly due to the notion that a person motivated by religion is unlikely to reason properly on historical matters. This assumption renders the works of religious Muslim scholars ineligible as valid secondary sources, relegating them to the status of primary sources. That is, the assumption is that works of Muslim scholarship may shed light on their authors and their milieus, but not on the historical questions with which they grapple. Referring to Nöldeke's work on the chronology of the Qurʾān, Fredrick Denny writes, "This type of scholarly-critical operation is not accepted by most Muslims, because it means treating the holy text just like any other text: generated by historical circumstances and understandable by means of historical-critical methodology" (Fredrick Denny, *An Introduction to Islam*, New York, MacMillan, 1985, p. 157). Denny is describing what he *expects* to be true of most Muslim scholars. Whatever the cause of Western scholars' occasional disengagement from the secondary scholarship produced by Muslims, the result is that at times they have missed key insights with regard to the history of the Qurʾān and, even more so, the *Ḥadīṯ*.

chosen randomly from among the *sūra*s that Bazargan does not break up into smaller blocks. The curves show the percentages of the verses with different verse lengths, where length is measured in the number of words.[29] For example, in *sūra* 77 almost 40 % of the verses have exactly three words. The different plots have a skewed bell shape, with a rapid rise followed by a slower decline. For each curve, Bazargan computes three parameters: mean verse length (MVL), mode, and height. Mean, or average, verse length is simply the total number of words divided by the total number of verses. The mode is the most frequent verse length. It is thus the verse length at which the curve achieves its peak. For example, in *sūra* 77, mode = 3. Finally, the height is simply the value of the peak, namely the percentage of verses that have the mode as their length.[30]

Bazargan observes that the mean, mode, and inverse of height tend to increase together in the Qurʾān, and often in like proportions. This gives him the idea to represent each block by the average of these three parameters in order to temper the aberrations of individual parameters—except that before taking the average, he makes the parameters comparable in size by dividing each by an appropriately chosen constant. Thus for each block he calculates a characteristic number given by $10 \times \frac{1}{3}\left(\frac{\text{mean}}{10} + \frac{\text{mode}}{8} + \frac{15}{\text{height}}\right)$ and reorders the blocks in order of increasing characteristic number to obtain his chronology. Table 1 shows the resulting chronology. This table lists and numbers the blocks in the chronological order—*e.g.* Block 2 came after Block 1. The notation "(2) 74: 1-7" means that "Block 2 is defined as verses 1-7 in *sūra* 74". Table 2 organizes the same data according to *sūra* number. In this table, "(2) 164: 40-152" means "*sūra* 2 contains Block 164, consisting of verses 40-152". The question remains as to how Bazargan determines the block divisions.

Bazargan leaves fifty-nine *sūra*s intact and divides the rest into smaller blocks. His doing so is consistent with the pre-modern and modern scholarly insight that *sūra*s may contain materials from different periods. In dividing the

---

[29] Bazargan does not count *lā, lam, law, hal*, or *yā* as distinct words, but he counts *wa-lā*, *fa-lā*, *a-lam*, *a-fa-lā, law lā, yā ayyuhā*, and *a-fa* as one word apiece (Bāzargān, *Sayr*, vol. I, p. 17-20 / 35-8 / 47). This may not be a standard way of defining words, but as Bazargan points out, for our purposes all that matters is that one be consistent in the way one counts. My own way of counting follows the usual definition of what a word is. This is reflected in all the statistics and plots I provide.

[30] Incidentally, I have found that results obtained by using the mode and height are too sensitive to the exact number of words, meaning that a slight change in the number of words in a verse could make a large difference in the mode or height. In future studies mode and height should be replaced with more stable markers that still capture what they are intended to get at.

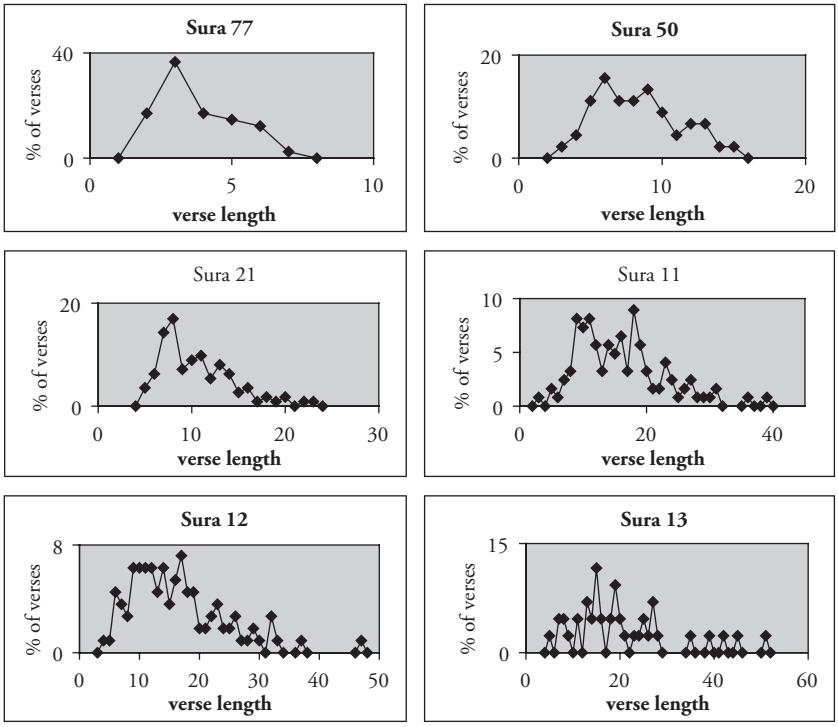

Figure 3. Characteristic curves for *sūra*s 77, 50, 21, 11, 12, 13, respectively Blocks 25, 87, 94, 118, 130, 169, listed in the chronological order according to Bazargan's scheme. The graphs show the percentage of verses with a given verse length, where verse length is measured in number of words. Note that as verse length increases from *sūra* to *sūra*, the peak tends to decline.

*sūra*s, Bazargan is guided by considerations of thematic unity, rhyme patterns, historical information, as well as verse length distribution, which he investigates by graphing characteristic curves for all *sūra*s and blocks. He observes, for example, that in many cases dividing a *sūra* as he does helps resolve incongruous characteristic curves into more typical-looking ones. In cases where he divides a *sūra* into more than two blocks, if two or more blocks in the *sūra* display similar verse length profiles, then because they belong to the same period in Bazargan's scheme, he combines them into a single block, even if they do not form a contiguous passage. The second volume of his book is devoted to discussing the block divisions for every *sūra*. Table 2 presents Bazargan's division of *sūra*s into blocks.

Table 1. ***Bazargan's Chronology. The blocks are numbered in the chronological order. A block is defined according to this format: (block number)* sūra*: verses, exclude verses. For example, "(1) 96: 1-5" means that Block 1 consists of verses 1-5 in* sūra *96. Verses marked "exclude" are excluded from calculations.***

(1) 96: 1-5. (2) 74: 1-7. (3) 103: 1-2. (4) 51: 1-6. (5) 102: 1-2. (6) 52: 1-8. (7) 112: 1-4. (8) 88: 1-5, 8-16. (9) 86: 11-17. (10) 82: 1-5. (11) 91: 1-10. (12) 108: 1-3. (13) 87: 1-7. (14) 85: 1-7, 12-22. (15) 81: 1-29. (16) 94: 1-8. (17) 93: 1-11. (18) 114: 1-6. (19) 79: 1-26. (20) 74: 8-10. (21) 92: 1-21. (22) 107: 1-7. (23) 70: 5-18. (24) 91: 11-15. (25) 77: 1-50, exclude 19, 24, 28, 34, 37, 40, 45, 47, 49. (26) 78: 1-36. (27) 74: 11-30, 32-56. (28) 106: 1-4. (29) 53: 1-22, 24-25. (30) 89: 1-14, 27-30. (31) 84: 1-25. (32) 80: 1-42. (33) 104: 1-9. (34) 109: 1-6, exclude 5. (35) 96: 6-19. (36) 88: 6-7, 17-26. (37) 75: 7-13, 20-40. (38) 95: 1-8, exclude 6. (39) 75: 1-6, 14-19. (40) 56: 1-96. (41) 55: 1-7, 10-27, 46-77, exclude 16, 18, 21, 23, 25, 47, 49, 51, 53, 55, 57, 59, 61, 63, 65, 67, 69, 71, 73, 75, 77. (42) 87: 8-19. (43) 1: 1-7. (44) 100: 1-11. (45) 69: 38-52. (46) 79: 27-46. (47) 111: 1-5. (48) 113: 1-5. (49) 90: 1-20. (50) 102: 3-8. (51) 105: 1-5. (52) 68: 1-16. (53) 89: 15-26. (54) 99: 1-8. (55) 86: 1-10, exclude 7. (56) 53: 33-62. (57) 101: 1-11. (58) 37: 1-182. (59) 82: 6-19. (60) 69: 1-3, 13-37. (61) 70: 19-35. (62) 83: 1-36. (63) 44: 43-59. (64) 23: 1-11. (65) 26: 52-227, exclude 127, 145, 164, 180, 67, 103, 121, 174, 190, 68, 104, 122, 140, 159, 175, 191, 110, 126, 131, 144, 150, 163, 179. (66) 38: 67-88. (67) 15: 1-5, 49-99. (68) 69: 4-12. (69) 97: 1-5. (70) 51: 7-60. (71) 54: 1-55, exclude 22, 32, 40, 21, 30. (72) 68: 17-52. (73) 44: 1-42. exclude 1. (74) 70: 1-4, 36-44. (75) 52: 9-20, 22-28. (76) 43: 66-80. (77) 71: 1-28. (78) 55: 28-45, 8-9, 78, exclude 28, 30, 32, 34, 36, 38, 40, 42, 45, 77-78, 8, 9. (79) 73: 1-19. (80) 20: 1-52, exclude 1. (81) 19: 75-98. (82) 52: 21, 29-49. (83) 15: 6-48. (84) 26: 1-51, exclude 1. (85) 76: 1-31. (86) 38: 1-25, 30-66. (87) 50: 1-45. (88) 36: 1-83, exclude 1. (89) 103: 3. (90) 23: 12-118. (91) 41: 1-8, exclude 1. (92) 43: 1-65, 81-89, exclude 1. (93) 33: 1-3, 7-8, 41-48, 63-68. (94) 21: 1-112. (95) 72: 1-28. (96) 78: 37-40. (97) 85: 8-11. (98) 19: 1-33, 41-74, exclude 1. (99) 98: 1-8. (100) 31: 1-11, exclude 1. (101) 30: 1-27, exclude 1. (102) 25: 1-77. (103) 20: 53-135. (104) 67: 1-30. (105) 14: 42-52. (106) 19: 34-40. (107) 16: 1-32, 41-64, 98-105, 120-128. (108) 18: 1-8, 60-110. (109) 32: 1-30, exclude 1. (110) 74: 31. (111) 17: 9-52, 61-65, 71-81, 101-111. (112) 40: 1-6, 51-60, exclude 1. (113) 2: 1-20, 153-157, 159-163, 204-209, 244-245, exclude 1. (114) 27: 1-93. (115) 39: 29-37, 53-66. (116) 45: 1-37, exclude 1. (117) 64: 1-18. (118) 11: 1-123. (119) 41: 9-36. (120) 30: 28-60. (121) 17: 1-8, 82-100. (122) 7: 59-155, 177-206. (123) 24: 46-57. (124) 22: 18-29, 42-69. (125) 6: 1-30, 74-82, 105-117. (126) 29: 1-69, exclude 1. (127) 34: 10-54. (128) 10: 71-109. (129) 38: 26-29. (130) 12: 1-111. (131) 28: 1-46, 85-88, 47-75, exclude 1. (132) 73: 20. (133) 40: 7-50, 61-85. (134) 53: 26-32, 23. (135) 18: 29-59. (136) 31: 12-34, exclude 15, 27. (137) 14: 1-5, 7-30, 32-41. (138) 42: 1-53, exclude 1. (139) 2: 30-39, 190-195. (140) 35: 4-7, 9-11, 13-17, 19-45. (141) 39: 1-28, 38-52. (142) 47: 1-38. (143) 8: 1-75. (144) 61: 1-14. (145) 41: 37-54. (146) 17: 53-60, 66-70. (147) 46: 1-14, 27-28, exclude 1. (148) 16: 33-40, 65-89, 106-119. (149) 5: 7-11, 20-26, 33-40. (150) 62: 1-11, exclude 3. (151) 3: 32-180. (152) 63:

Table 1 (*cont.*)

1-11. (153) 22: 1-17, 30-41, 70-78. (154) 3: 1-31, 181-200, exclude 1. (155) 7: 1-58, 156-176, exclude 1. (156) 59: 1-24, exclude 7, 15, 21-24. (157) 39: 67-75. (158) 34: 1-9. (159) 9: 38-70. (160) 10: 1-70. (161) 57: 1-29. (162) 16: 90-97. (163) 24: 1-34. (164) 2: 40-152. (165) 33: 4-6, 9-40, 49-52, 56-62, 69-73. (166) 4: 44-57, 131-175. (167) 6: 31-73, 83-104, 118-134, 154-165. (168) 13: 1-43. (169) 9: 71-129. (170) 48: 1-29. (171) 65: 8-12. (172) 5: 51-86. (173) 49: 1-18. (174) 28: 76-84. (175) 4: 1-43, 58-126. (176) 18: 9-28. (177) 9: 1-37. (178) 46: 15-26, 29-35. (179) 110: 1-3. (180) 14: 6, 31. (181) 58: 1-22. (182) 5: 27-32, 87-120. (183) 2: 21-29, 158, 165-189, 196-203, 210-242, 254, 261-283. (184) 60: 1-13. (185) 35: 1-3, 8, 12, 18. (186) 66: 1-12. (187) 6: 135-153. (188) 65: 1-7. (189) 4: 127-130, 176. (190) 24: 35-45, 58-64. (191) 5: 12-19, 41-50. (192) 2: 164, 243, 246-253, 255-260, 284-286. (193) 33: 53-55. (194) 5: 1-6.

Table 2. ***The division of the* sūra*s into blocks. The information is sorted by* sūra *number. The format is* (sūra) *block: verses, exclude verses. For example, "(1) 43: 1-7" means that* sūra *1 contains Block 43, which covers verses 1-7. Verses marked "exclude" are excluded from calculations.***

(1) 43: 1-7. (2) 113: 1-20, 153-157, 159-163, 204-209, 244-245, exclude 1. (2) 183: 21-29, 158, 165-189, 196-203, 210-242, 254, 261-283. (2) 139: 30-39, 190-195. (2) 164: 40-152. (2) 192: 164, 243, 246-253, 255-260, 284-286. (3) 154: 1-31, 181-200, exclude 1. (3) 151: 32-180. (4) 175: 1-43, 58-126. (4) 166: 44-57, 131-175. (4) 189: 127-130, 176. (5) 194: 1-6. (5) 149: 7-11, 20-26, 33-40. (5) 191: 12-19, 41-50. (5) 182: 27-32, 87-120. (5) 172: 51-86. (6) 125: 1-30, 74-82, 105-117. (6) 187: 135-153. (6) 167: 31-73, 83-104, 118-134, 154-165. (7) 155: 1-58, 156-176, exclude 1. (7) 122: 59-155, 177-206. (8) 143: 1-75. (9) 177: 1-37. (9) 159: 38-70. (9) 169: 71-129. (10) 160: 1-70. (10) 128: 71-109. (11) 118: 1-123. (12) 130: 1-111. (13) 168: 1-43. (14) 137: 1-5, 7-30, 32-41. (14) 180: 6, 31. (14) 105: 42-52. (15) 67: 1-5, 49-99. (15) 83: 6-48. (16) 107: 1-32, 41-64, 98-105, 120-128. (16) 148: 33-40, 65-89, 106-119. (16) 162: 90-97. (17) 121: 1-8, 82-100. (17) 111: 9-52, 61-65, 71-81, 101-111. (17) 146: 53-60, 66-70. (18) 108: 1-8, 60-110. (18) 176: 9-28. (18) 135: 29-59. (19) 98: 1-33, 41-74, exclude 1. (19) 106: 34-40. (19) 81: 75-98. (20) 80: 1-52, exclude 1. (20) 103: 53-135. (21) 94: 1-112. (22) 153: 1-17, 30-41, 70-78. (22) 124: 18-29, 42-69. (23) 64: 1-11. (23) 90: 12-118. (24) 163: 1-34. (24) 190: 35-45, 58-64. (24) 123: 46-57. (25) 102: 1-77. (26) 84: 1-51, exclude 1. (26) 65: 52-227, exclude 127, 145, 164, 180, 67, 103, 121, 174, 190, 68, 104, 122, 140, 159, 175, 191, 110, 126, 131, 144, 150, 163, 179. (27) 114: 1-93. (28) 131: 1-46, 85-88, 47-75, exclude 1. (28) 174: 76-84. (29) 126: 1-69, exclude 1. (30) 101: 1-27. exclude 1. (30) 120: 28-60. (31) 100: 1-11, exclude 1. (31) 136: 12-34, exclude 15, 27. (32) 109: 1-30, exclude 1. (33) 93: 1-3, 7-8, 41-48, 63-68. (33) 165: 4-6, 9-40, 49-52, 56-62, 69-73. (33) 193: 53-55. (34) 158: 1-9. (34) 127: 10-54. (35) 185: 1-3, 8, 12, 18. (35) 140: 4-7, 9-11, 13-17, 19-45. (36) 88: 1-83, exclude 1. (37) 58: 1-182. (38) 86: 1-25, 30-66. (38) 129: 26-29. (38) 66: 67-88. (39) 141: 1-28, 38-52. (39) 115: 29-37, 53-66. (39) 157:

Table 2 (*cont.*)

67-75. (40) 112: 1-6, 51-60, exclude 1. (40) 133: 7-50, 61-85. (41) 91: 1-8, exclude 1. (41) 119: 9-36. (41) 145: 37-54. (42) 138: 1-53, exclude 1. (43) 92: 1-65, 81-89, exclude 1. (43) 76: 66-80. (44) 73: 1-42, exclude 1. (44) 63: 43-59. (45) 116: 1-37, exclude 1. (46) 147: 1-14, 27-28, exclude 1. (46) 178: 15-26, 29-35. (47) 142: 1-38. (48) 170: 1-29. (49) 173: 1-18. (50) 87: 1-45. (51) 4: 1-6. (51) 70: 7-60. (52) 6: 1-8. (52) 75: 9-20, 22-28. (52) 82: 21, 29-49. (53) 29: 1-22, 24-25. (53) 134: 23, 26-32. (53) 56: 33-62. (54) 71: 1-55, exclude 22, 32, 40, 21, 30. (55) 41: 1-7, 10-27, 46-77, exclude 16, 18, 21, 23, 25, 47, 49, 51, 53, 55, 57, 59, 61, 63, 65, 67, 69, 71, 73, 75, 77. (55) 78: 28-45, 8-9, 78, exclude 28, 30, 32, 34, 36, 38, 40, 42, 45, 78, 8, 9. (56) 40: 1-96. (57) 161: 1-29. (58) 181: 1-22. (59) 156: 1-24, exclude 7, 15, 21-24. (60) 184: 1-13. (61) 144: 1-14. (62) 150: 1-11, exclude 3. (63) 152: 1-11. (64) 117: 1-18. (65) 188: 1-7. (65) 171: 8-12. (66) 186: 1-12. (67) 104: 1-30. (68) 52: 1-16. (68) 72: 17-52. (69) 60: 1-3, 13-37. (69) 68: 4-12. (69) 45: 38-52. (70) 74: 1-4, 36-44. (70) 23: 5-18. (70) 61: 19-35. (71) 77: 1-28. (72) 95: 1-28. (73) 79: 1-19. (73) 132: 20. (74) 2: 1-7. (74) 20: 8-10. (74) 27: 11-30, 32-56. (74) 110: 31. (75) 39: 1-6, 14-19. (75) 37: 7-13, 20-40. (76) 85: 1-31. (77) 25: 1-50, exclude 19, 24, 28, 34, 37, 40, 45, 47, 49. (78) 26: 1-36. (78) 96: 37-40. (79) 19: 1-26. (79) 46: 27-46. (80) 32: 1-42. (81) 15: 1-29. (82) 10: 1-5. (82) 59: 6-19. (83) 62: 1-36. (84) 31: 1-25. (85) 14: 1-7, 12-22. (85) 97: 8-11. (86) 55: 1-10, exclude 7. (86) 9: 11-17. (87) 13: 1-7. (87) 42: 8-19. (88) 8: 1-5, 8-16. (88) 36: 6-7, 17-26. (89) 30: 1-14, 27-30. (89) 53: 15-26. (90) 49: 1-20. (91) 11: 1-10. (91) 24: 11-15. (92) 21: 1-21. (93) 17: 1-11. (94) 16: 1-8. (95) 38: 1-8, exclude 6. (96) 1: 1-5. (96) 35: 6-19. (97) 69: 1-5. (98) 99: 1-8. (99) 54: 1-8. (100) 44: 1-11. (101) 57: 1-11. (102) 5: 1-2. (102) 50: 3-8. (103) 89: 3. (103) 3: 1-2. (104) 33: 1-9. (105) 51: 1-5. (106) 28: 1-4. (107) 22: 1-7. (108) 12: 1-3. (109) 34: 1-6, exclude 5. (110) 179: 1-3. (111) 47: 1-5. (112) 7: 1-4. (113) 48: 1-5. (114) 18: 1-6.

In rearranging the blocks, Bazargan follows his ordering method strictly, but makes an exception for the first five verses (twenty words) of *sūra* 96, which he makes the first block.[31] In fact, their proper location is somewhat later, between blocks 36 and 37, although one may also join these verses with the rest of the *sūra*, namely with what is now Block 35. In my own investigations, I have strictly preserved Bazargan's ordering with this one exception.

I will now describe my decisions about what to count. These decisions are mostly about minor things that do not affect the overall results, but it's useful to specify them to make it possible for other researchers to reproduce my results. I have excluded certain verses from calculations, as indicated in Table 1 and Table 2. These verses fall into three categories. First, Bazargan counts only once verses that are repeated to serve as a refrain. For example, the sentence *fa-bi-ayyi ālāʾi rabbikumā tukaḏḏibān* occurs thirty-one times in *sūra* 55,

[31] Bazargan deferred to some historical reports that identify the beginning verses of *sūra* 96 as the first revelation.

but Bazargan excludes the repetitions, observing that doing so yields a more typical-looking characteristic curve. I too have excluded the refrains, counting only their first occurrences.[32] Second, Bazargan excludes some verses from his calculations due to thematic and stylistic considerations without assigning these excluded verses to any particular block.[33] He also excludes twelve words from Kor 60, 4 (Block 184). In my own calculations, I adopt Bazargan's exclusions except, merely for ease of computation, in this last case. One may argue with some of Bazargan's choices, but due to the small number of such cases, such disputation would be immaterial to my results. Third, I have removed the mysterious "detached letters" from computation. Where these letters take up entire verses of their own, I have excluded the verses.[34] Where the mysterious letters form only part of a verse, I have left aside the letters while including the remainder of the verse.[35] Such details will not change my results, concerned as they are with broad patterns.

There are various systems of dividing up the text into verses. They differ among themselves typically by one, two, or a few verses per *sūra*.[36] Bazargan's numbering of verses follows a particular Qur'ān published in Iran in HS 1328/1949 to which I do not have access.[37] Examining his citations, its numbering appears to agree with that of the Flügel edition through *sūra* 79, after which it switches to another system. I have gone through his citations and converted them into the numbering system of the Egyptian standard edition, *i.e.* the Kūfan system, which is what all my tables and references reflect.

A few of the smaller blocks are highly sensitive to the uncertainties in verse division. For example, Block 110 (*sūra* 74: 31) is three verses in Bazargan's reckoning and only one in mine, and Block 179 (*sūra* 110: 1-3) is one verse in his reckoning and three in mine. That translates into a three-fold difference in mean verse length. These passages, however, are small, and their placement

---

[32] Here are the repeated verses I have excluded: in *sūra* 26: verses 127, 145, 164, 180, 67, 103, 121, 174, 190, 68, 104, 122, 140, 159, 175, 191, 110, 126, 131, 144, 150, 163, 179; in *sūra* 54: verses 22, 32, 40, 21, 30; in *sūra* 55: verses 16, 18, 21, 23, 25, 28, 30, 32, 34, 36, 38, 40, 42, 45, 47, 49, 51, 53, 55, 57, 59, 61, 63, 65, 67, 69, 71, 73, 75, 77; in *sūra* 77: verses 19, 24, 28, 34, 37, 40, 45, 47, 49; in *sūra* 109: verse 5.

[33] Here is a list: in *sūra* 31: verses 15, 27 (Block 136); in *sūra* 55: verses 8-9, 78 (Block 78); in *sūra* 59: verses 7, 15, 21-24 (Block 156); in *sūra* 62: verse 3 (Block 150); in *sūra* 86: verse 7 (Block 9); in *sūra* 95: verse 6 (Block 38).

[34] These cases occur in *sūra*s 2, 3, 7, 19, 20, 26, 28, 29, 30, 31, 32, 36, 40, 41, 42, 43, 44, 45, 46.

[35] These cases include *sūra*s 10, 11, 12, 13, 14, 15, 27, 38, 50, 68.

[36] Anton Spitaler, *Die Verszählung des Koran nach islamischer Überlieferung*, Munich, Verlag der Bayerischen Akademie der Wissenschaften, 1935.

[37] Bazargan cites it as "Qur'ān in the handwriting of Mr. Ḫušnavīs, Tehran, Tīr 1328, Kitābfurūšī-i Islāmiyya" (Bāzargān, *Sayr*, vol. I, p. 14 / 32 / 39).

makes little difference in the results of this essay, which are broad-based, concerned as they are with average properties of large groups of text. Conversely, the methods used in this article are ill-suited to placing blocks of such small size in the chronological sequence.

The fact that semantic and thematic features, in addition to stylistic ones, play an important role in the process of segmentation of *sūra*s into blocks may elicit the objection that non-stylistic criteria are applied in a task that was supposed to be free of them.[38] One may address this concern by noting that the ultimate determinant of the chronology is the reordering procedure, which is purely style-based. This reordering procedure can help mitigate mistakes made in the process of segmentation. Should Bazargan mistakenly have divided passages that really belong together, the reordering procedure, if it is sound, should assign them to about the same period. However, this does not mean that mistaken divisions will be without a cost. Once divided, the text becomes smaller, and its analysis more vulnerable to sampling error, leading to loss of precision in reordering.[39] The number of phases displaying concurrent smoothness may be reduced, leading one to confirm less of the chronological sequence than one might do otherwise. Loss of precision, however, is not loss of accuracy. "Stanford University is on earth" may be less precise than "Stanford University is in California", but it is no less accurate. In sum, reliance on meaning at the stage of segmentation does not fundamentally prejudice the final chronology nor makes it less accurate, even though it may entail loss of information and make it less precise.

I have not discussed whether the principles behind Bazargan's proposed chronological sequence are sound. The reason why is that the Criterion of Concurrent Smoothness shall be the judge of the sequence of passages that result from them. *How* Bazargan arrived at his chronology is immaterial as long as it exhibits concurrent smoothness. In fact, if one proposed a chronology based on a purely random procedure, and if this chronology happened to yield greater concurrent smoothness, then it would be a superior proposal.

If a proposed sequence achieves concurrent smoothness, others that are broadly similar to it may do so as well. There may be various proposed chronologies with large numbers of small differences which achieve a similar degree of concurrent smoothness, and these will be considered as equally corroborated. This is because the corroboration offered by statistical methods involves average characteristics of long texts or large aggregates of short texts, not small units of text in isolation. This is one reason why readers must resist the temptation to take either Bazargan's chronological list or my own recasting of it in a more precise way than they are intended. Where I identify one group of texts

[38] See "Motivating the Approach in this Essay", above in Section 1.

[39] For sampling error, see "Stylometry Demystified", above in Section 1.

as having come after another, this claim holds only in an average sense; it does not mean that every text in one group came after every text in the other group. Increasing precision is a long-term goal, and this essay represents only the beginning of the journey.

The definition of "groups"

It is convenient (and in the case of small blocks necessary) to combine the blocks into larger groups. This has the advantage of simplifying graphs and presentations of results. The manner of aggregating the blocks into groups is arbitrary as long as consecutive blocks, which have similar verse-length profiles, are grouped together. I find it convenient to combine the 194 blocks into twenty-two groups corresponding in a rather approximate fashion to the twenty-three years of revelation as defined by Bazargan, which is not to say that I am committed to his reckoning of years. Even he noted that the true date of a passage may be two or three years off its assigned date. I will speak loosely of these as "Bazargan's groups". For each group, Table 3 lists the blocks it contains and the number of words in it. It should be noted that the first group is rather small, which entails a larger sampling error, warranting caution about whether it can be characterized adequately by the chosen markers of style. I will come back to this point in the Conclusion. As each group includes a set of adjacent blocks, verse length tends to increase from one group to the next. I now proceed to examine how different markers of style behave over these groups.

Table 3. ***The Groups Defined. The block numbers are those defined in Table 1. Each group contains passages that belong to the same period according to Bazargan. Our task is to see how style varies over these consecutive groups.***

Group	Blocks	Words	Group	Blocks	Words	Group	Blocks	Words
1	2-16	415	9	117-121	3494	17	161-164	3504
2	17-34	1256	10	122-126	4463	18	165-167	3860
3	1, 35-65	3916	11	127-131	4431	19	168-174	3959
4	66-82	3214	12	132-138	3660	20	175-178	4018
5	83-91	3624	13	139-143	3423	21	179-183	4019
6	92-101	3676	14	144-150	2338	22	184-194	3766
7	102-110	4430	15	151-154	4389			
8	111-116	3479	16	155-160	3920			

3. Univariate Assessments of Smoothness

Univariate markers of style are those that involve just one variable, thus representing the twenty-two groups by one number apiece. Variables considered here include mean verse length, mean word length, the standard deviation of word length, and the frequency of hapax legomena. A key question is whether these markers display smooth behavior over Bazargan's twenty-two groups. That is, do groups that are near each other in Bazargan's sequence tend to have similar stylistic profiles? The extent to which different, independent markers of style vary in a smooth fashion over a sequence is a measure of concurrent smoothness, hence of the degree of confirmation of the sequence as the true chronological one.

Univariate Marker of Style: Mean Verse Length (MVL)

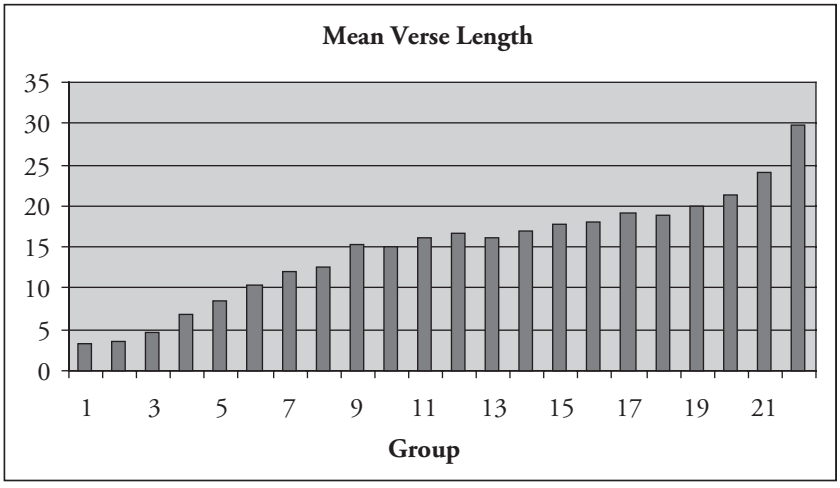

Figure 4. Mean Verse Length (MVL) vs. Group number. The height of the third column, for example, gives the size of MVL for the third group. It can be seen that in Bazargan's sequence, average verse length varies relatively smoothly over time.

Figure 4 shows the mean (*i.e.* average) verse length (MVL) of each group, where the length of a verse is the number of words in it. Each column represents a group, and its height the size of the MVL of that group. MVL tends to increase over Bazargan's sequence. This is to be expected, as his sequence was devised precisely to achieve this effect: he posited that verse length increased gradually over time.

One cannot yet claim concurrent smoothness, since I have not yet checked whether markers of style other than verse length also vary smoothly. But one may make a number of observations. First, note the great range in the variation of verse length. Between the group with the shortest verses and that with the longest, the difference in mean length is greater than a factor of nine! This huge difference hints that verse length might be an effective discriminator of style. Second, the variation in verse length is not discrete in nature: there is a continuum between the extremes of MVL. This is not a trivial point, as one would have had no reason to particularly expect this in advance. This suggests that the hypothesis of gradual change is worth examining. Third, *if* it is true that verse length is an effective indicator of relative time, one should expect lesser discriminatory efficacy over Groups 9-19, which exhibit significantly less variation. Since these groups have relatively similar verse lengths, it becomes easier to imagine them occurring in a different order than shown. For example, if one were to switch Groups 11 and 12, that would make a smaller dent in smoothness than if one switched Groups 4 and 5. This undermines confidence in the accuracy of the second half of Bazargan's proposed sequence. All of the above hunches will be confirmed more rigorously in later sections using multivariate markers of style.

One way to examine the efficacy of MVL as a criterion is to see what it does to passages that we already suspect belong to the same period. This is because MVL is effective only to the extent that passages from the same period have similar mean verse lengths. The simplest way to check this is to take a relatively coherent passage, bisect it, and see if the two halves have similar MVLs. To that end, I consider all *sūra*s which Bazargan leaves intact and which have 570 or more words. There are twelve such *sūra*s, and they are shown in Table 4.

I have divided each text in Table 4 into two nearly equal parts, which I loosely call "halves". This table and Figure 5 give the MVL for each half-text. The texts are listed in order of increasing block number, hence approximately increasing MVL. The *distance between two halves* is defined as the difference between their MVLs. In the topmost graph, the length of each line segment represents the distance between two halves. It is evident that, on the whole, the two halves of the texts have similar MVLs. This similarity is greatest, however, for the "earlier" ones, which have shorter verses.[40]

[40] This remains true if one measures dissimilarity as the percentage of the difference relative to the MVL, *i.e.* as the distance between the two halves divided by the MVL.

Table 4. ***Twelve intact* sūra*s, arranged in order of increasing block number. Each* sūra *is divided into two halves with almost the same number of words. The table gives the MVL for each half, the distance between each half (defined as the absolute value of the difference of their MVLs), and clustering quality.***

Text number	*Sūra* (block)	Number of words	MVL in 1st half	MVL in 2nd half	Distance betwn. the two halves	Clustering Quality (m=3)
1	37 (058)	865	4.97	4.56	0.41	16.7
2	36 (088)	729	8.71	9.08	0.36	9.72
3	21 (094)	1174	10.7	10.2	0.52	5.42
4	25 (102)	896	12.2	11.1	1.12	2.09
5	27 (114)	1158	12.9	12.0	0.89	2.63
6	11 (118)	1945	17.0	14.8	2.21	0.63
7	29 (126)	977	14.6	14.1	0.50	4.02
8	12 (130)	1793	16.2	16.1	0.06	24.7
9	42 (138)	861	19.0	14.8	4.15	0.36
10	08 (143)	1243	15.3	18.1	2.77	0.51
11	57 (161)	574	18.3	21.7	3.44	0.83
12	13 (168)	853	21.1	18.7	2.36	1.19

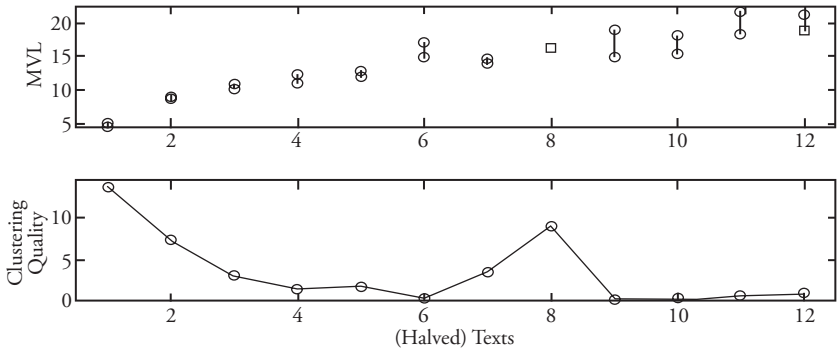

Figure 5. Pair distances and Clustering Quality. *Top*: MVLs for each pair of halves (vertical axis) vs. text number (horizontal axis). The halves of the first eight texts have nearer MVLs than do the halves of the last four texts. *Bottom*: Clustering Quality (vertical axis) vs. text number (horizontal axis). Clustering Quality for a given text indicates how easily one may recognize its two halves as belonging together. The larger the Clustering Quality is, the easier that task. There is a decreasing trend, thus clustering is more successful for the earlier blocks.

In the topmost graph in Figure 5, I use our prior knowledge of which halves belong together to examine how close the halves are. Now imagine the reverse of this process. Suppose you were given the MVL values of all twenty-four halves, but were not told which half goes with which, and were asked to guess the correct pairings. How successful would you be? You would form the correct clusters in five cases (texts 1, 2, 3, 7, and 8), but probably fail in the other seven cases.

Given a text, two factors contribute to one's success or failure in assigning its halves to the same cluster: how close they are to each other, and how far apart they are from the halves of other texts. The farther apart from other pairs, and the closer together they are, the greater one's ability to correctly join the halves. Thus one may define *Clustering Quality* for a given pair of halves as: the mean of distances of the given pair from, say, the three nearest pairs (m =3) divided by the distance between the two halves of the given pair. (Here, the distance between any two pairs of halves is defined as the distance between their respective midpoints.)[41] This quantity signifies the ease with which one can successfully assign the two halves of a given pair together. Figure 5 (bottom graph) gives a plot of Clustering Quality for each pair of halves, while Table 4 lists its values.

The test confirms the validity of MVL as a marker of style that can provide information about chronology. For, as seen above, texts that we know already date from the same time tend to have similar MVLs. Such correspondence is very striking for passages with shorter verses. The test, however, also suggests that the discriminatory effectiveness of MVL may decline with increasing MVL. If this is true—and more tests are needed—then given that Bazargan's ordering is based entirely on verse length, confidence in its later parts is less justified.

The possibly better performance of MVL with passages with shorter verses is perhaps due to two facts. First, as already noted, MVL appears to vary more rapidly over the "earlier" groups. Second, in earlier groups verse length tends to be more consistent. That is, within each group, verses tend to have lengths that are close the MVL. By contrast, verses in "later" passages display larger deviations from the mean verse length.

Figure 6 shows graphs of standard deviation, which is a measure of the average amount of deviation of the verse lengths from the mean verse length, MVL.

---

[41] For example, suppose the first two halves have 1 and 2 as their average verse lengths respectively, and the next two halves have 3 and 4 as that value. Then the midpoint is 1.5 for the first

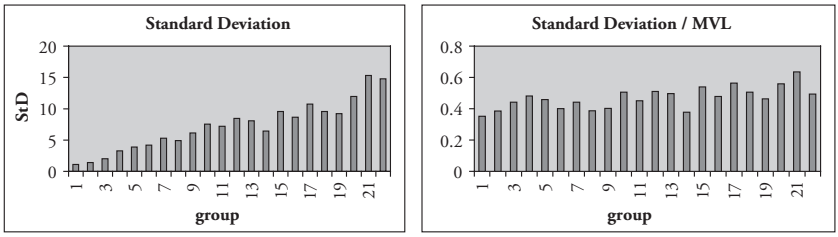

Figure 6. Left: Standard deviation of verse length indicates how much, on average, verse length differs from the MVL. Passages with longer verses display a greater amount of spread around the mean verse length. *Right*: Dividing the standard deviation of verse length by MVL gives the percentage by which verse length deviates from MVL. The deviation is about 40% in the earlier groups and 50% in the later ones.

Univariate Markers of Style: Word Length Mean and Standard Deviation

By the *length* of a word, I mean the number of consonants and vowels it contains.[42] Figure 7 shows the mean and standard deviation of word length for each group. The mean word length is simply the average word length, which is obtained by summing the word lengths and dividing by the number of words. The standard deviation is a measure of how much, on average, word length differs from the mean word length. It indicates whether word length tends to stay at about the same value. A small standard deviation means that word lengths tend to stay close to the mean, while a large value indicates greater diversity in word length. (It is obtained by subtracting the length of each word from the mean, squaring this quantity, obtaining the average of these squares for all the words, and then taking the square root. In other words, it is the square root of the mean of the squares.)

As shown in the figure, mean length shows very little variation compared to standard deviation. Mean word length ranges from about 6.3 to 6.7, a variation of only 6 %. On the other hand, standard deviation ranges from 0.54 to

pair, and 3.5 for the second. Subtracting these, one obtains 2 as the distance between the first pair of halves and the second pair of halves.

[42] I count vowels that are not written. I do not count elided *alif*s. I count the *lām* in the definite particle regardless of whether what follows it is a sun letter or moon letter. I count consonants with *šadda* twice. As usual, the exact way one defines a thing is unimportant. What matters is that one is consistent in applying whatever definition one has picked.

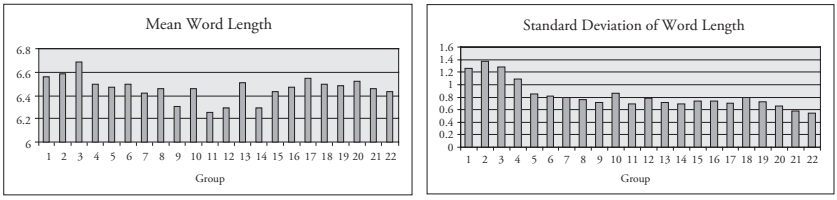

Figure 7. Mean word length (left) shows only slight variation, making the shape of the curve more vulnerable to sampling error. By contrast, there is a significant variation in the standard deviation (right), making the differences more meaningful and hinting at genuine variation in style.

1.37, a variation of 154 %. This indicates that the shape of the graph of standard deviation is less vulnerable to the random fluctuations of sampling error, as these oscillations are drowned out by the larger trends. Thus the pattern captures a genuine variation in style, not just random fluctuations arising from sampling error.

The graph of the mean is perhaps slightly smoother than if the blocks were arranged at random. There is only a very vague tendency for groups of similar mean lengths to be grouped together. One cannot conclude much.

A variable varies "smoothly" if groups that are located near each other in the sequence have similar values. The more this holds, the smoother a graph. In a smooth variation, if one takes the differences in values of consecutive groups and adds them up, this sum will be smaller than what one would obtain on average from a random arrangement of the groups. Visually, this translates into a graph with fewer jags or smaller jags. And the more there are local or global trends, the greater the amount of smoothness locally or globally.

The graph of standard deviation displays fairly smooth behavior in its left side. After that, there is no general smoothness except at the very end, although one may discern the outline of an overall trend. But this lack of smoothness in the middle part might be a sign of the fact that the groups have about the same height, so sampling error may be the cause of the jaggedness. In addition, note that although the internal chronology of Groups 6-22 remains in doubt, at least it is clear that as a whole they belong to the right side of Groups 1-5; if they were moved to the left of Group 1, that would create a severe discontinuity.

This graph hints at the possible validity of the first quarter of Bazargan's chronology, for it means that two independent markers of style vary in a smooth fashion, the markers being (1) mean verse length and (2) the standard

deviation of word length. The graph also suggests that Groups 21 and 22 belong together.

Univariate Marker of Style: Hapax Legomena

I use the term "hapax legomena" a bit loosely. It properly refers to words that occur only once in a corpus. Here, I use it to refer to morphemes that occur just once in the entire Qurʾān. A morpheme can be a word *or part of a word*. So, for example, if the term *ǧamīl* occurs here with a definite article (*l-ǧamīl-u*) and there with an indefinite accusative case ending (*ǧamīl-an*), then I count these as two occurrences of the same morpheme, not as two different morphemes.

I use the transliteration of the Qurʾān developed by Rafael Talmon and Shuly Wintner, which uses hyphens to divide words into morphemes.[43] Do they divide words in the right way? Actually, there is no one correct way of dividing words. The chief requirement is that the division be done in a consistent manner, and the Wintner-Talmon transliteration meets this requirement.

There are about 4,000 hapax legomena in the Qurʾān, accounting for slightly over half of the total number of distinct morphemes. These, however, are not distributed in the groups in an even fashion. To get a sense of their distribution, I have, for each group, added up the total number of hapax legomena and divided this number by the total number of words, yielding the percentage of words that have a hapax legomenon. The resulting percentages, graphed in Figure 8, provide a measure of vocabulary richness.

The graph is striking. The amount of variation represents a six-fold difference over its range. One gets a sense of rapid stylistic change in the left side towards a steady state, which is reached by Group 5. Thus, in the left side of the graph one observes not only smoothness, but also a trend. As for the

[43] See Rafi Talmon and Shuly Wintner, "Morphological Tagging of the Qurʾān", in *Proceedings of the Workshop on Finite-State Methods in Natural Language Processing, an EACL'03 Workshop*, Budapest, Hungary, April 2003; and Judith Dror, Dudu Shaharabani, Rafi Talmon, and Shuly Wintner, "Morphological analysis of the Qurʾān", *Literary and Linguistic Computing*, 19/4 (2004), p. 431-52. For the on-line tagged Qurʾān developed by Rafi Talmon and Shuly Wintner (and their students), see http://cl.haifa.ac.il/projects/quran/. One may use the program to search for morphemes as in the following example. To search for the morpheme "*wa*", type the following command in the spot for SQL expressions in the upper right part of the program's window, and then click on "Analyze": "SELECT DISTINCT tbl0.location , tbl0.Word , tbl0. full_analyse FROM qortbl2 AS tbl0 WHERE (tbl0.Word LIKE "%-wa" OR tbl0.Word LIKE "%-wa-%" OR tbl0.Word LIKE "wa-%" OR tbl0.Word LIKE "wa") ORDER BY location". To search for other morphemes, replace all occurences of "wa" in this expression with the morpheme of interest.

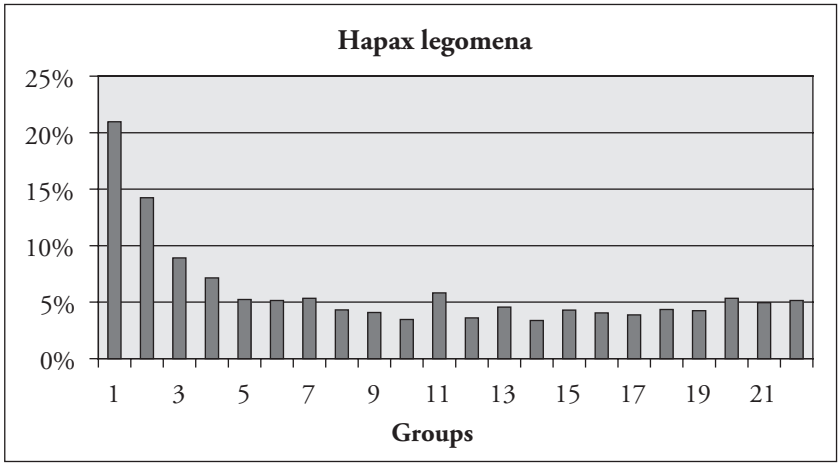

Figure 8. Relative frequencies of hapax legomena, *i.e.* the fraction of words in each group that have a hapax legomenon, obtained by dividing the total number of hapax legomena in each group by the number of words in the group.

remainder of the groups, one can conclude nothing about their correct sequence other than observing that they belong together on the right side. If they were moved to the left of Group 1, a great discontinuity would appear.

Univariate Markers of Style: Summary of Results

We have seen graphs of four independent indicators of style: mean verse length, mean word length, standard deviation of word length, and hapax legomena. In these graphs, except perhaps for mean word length, one observes fairly smooth behavior over the first five or so groups. This speaks in favor of the proposition that these groups are arranged in the true chronological sequence (or reverse-chronological sequence). With even greater justification, one can say that the remaining groups collectively belong to the right side of this sequence, since if one moved them to the left side of Group 1, then that would create a drastic discontinuity in three of the graphs. However, so far one cannot say much about the ordering of Groups 6-22 relative to one another. Multivariate methods of analysis, which are more effective, will shed more light on that part of the sequence.

4. A Non-Technical Introduction to Multivariate Methods

The most powerful stylometric methods employ many variables simultaneously. For example, in a later section I represent each of Bazargan's groups by a row of twenty-eight numbers consisting of the relative frequency counts of twenty-eight morphemes. The goal will be as before: to assess the similarity of nearby groups in Bazargan's sequence in order to check for smoothness. But while such a task is easy for univariate markers of style such as word length, it is more difficult to see how to compare texts that are represented by twenty-eight numbers apiece. One may do so using the techniques of *multivariate analysis*.

In the present section, I will illustrate two multivariate techniques, PCA and MDS, by using the frequency counts of only *three* morphemes to represent each group. This allows pictorial representation of the groups as points in a three-dimensional space, making it easier to grasp the concepts. When applied to this three-dimensional dataset, the PCA and MDS techniques generate a two-dimensional (flat-surface) diagram that represents the dissimilarity of two texts by the distance between them. Thus, two texts that are stylistically similar in that they have similar morpheme-frequency profiles are placed near each other. The resulting diagram can be used to determine whether the progression over the sequence of Bazargan's groups is smooth or jagged.

Figure 9 shows the relative frequency counts of three morphemes per group, namely *llāh* (marked as "llaah") and the indefinite case endings *an* and *in*. The relative frequency count of *llāh* in Group 2, for example, is the percentage of morphemes in Group 2 that are *llāh*. It is obtained by taking the total number of occurrences of *llāh* in Group 2 and dividing it by the total number of occurrences of morphemes in that group. Each group is thus represented with a vector, *i.e.* a row of numbers: in this case three numbers. To assess the smoothness of the transitions, one needs to discern which groups are similar to which, since smoothness means that consecutive groups have similar profiles. You can see that Groups 1 and 3 have very similar profiles, but an aberrant Group 2 interposes between them. Other similar pairs include, for example, Groups 4 and 5, 6 and 7, and 21 and 22.

The similarity or dissimilarity of two groups can be expressed through the notion of the *distance* between them, which can be defined in terms of the difference between their respective frequency count vectors. Suppose you wish to assess the similarity of Groups 1 and 3. To quantify their dissimilarity, one first calculates the difference in height of the *llāh*-columns of the two groups—by subtracting the height of the shorter column of *llāh* from the taller one. One then does the same for the pair of *an*-columns and the pair of *in*-columns.

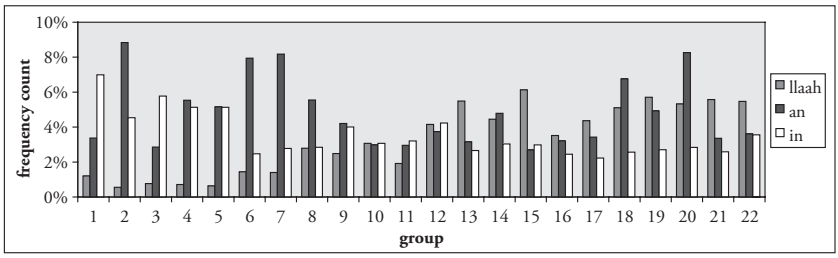

Figure 9. Group profiles in terms of the relative frequency counts of three morphemes.

One then adds up these differences to get a sense of the overall difference. This calculation gives what is called *the city-block distance* between the two groups, which is the notion of distance used in this essay.[44] Obviously, similar texts have a small city-block distance.

A crucial point is that just as, in the last paragraph, I calculated distances between pairs of groups represented in three dimensions, one may also do so in higher dimensions. That is, if one represents each group by the frequency counts of twenty-eight morphemes, one may calculate the city-block distances between any two groups just as before. The difference is that we can readily visualize data in a two- or three-dimensional space, whereas higher dimensions require special techniques to enable visualization.

Considering that each group is represented by three frequency counts, one may visualize the groups as points in a three-dimensional space, where each axis represents the frequency count of one morpheme. Figure 10 provides visualizations of the same set of points from two different angles. Labeling each group would make the images too cluttered, so I have opted for connecting consecutive groups with straight lines. The topmost point is Group 1.

The city-block distance between two points is the length of the shortest path between them if one were allowed to move only in lines parallel to the axes.[45] Observe that Groups 1 and 3 are close to each other as expected, but in the path from 1 to 3, the aberrant Group 2 has caused a jag along the direction

[44] Alternatively, instead of simply adding up the differences in height, one may first square these differences, then add them up, and then take the square root of the sum. This gives what is known as the Euclidean distance or the l_2 distance. This is an equally popular measure of distance, though not the one used in this essay. The results in this article are conservative and robust enough to be unaffected by the choice of the distance measure.

[45] On the other hand, the Euclidean distance between two groups (as defined in the last footnote) would be simply the length of the straight line connecting them.

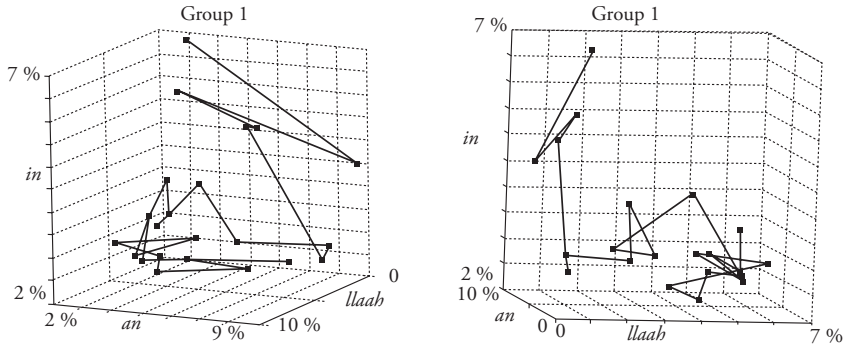

Figure 10. Bazargan's groups represented as points in three-dimensional space. The two images depict the same points, but from different angles. Group 1 is the topmost point. Consecutive groups are connected with straight lines.

of the *an* axis. By contrast, the path from Group 7 to Group 9 is smooth, with Group 8 in a nice, intermediate position.

As mentioned before, data in a twenty-eight-dimensional space cannot be visualized in the way three-dimensional data can. However, there are multivariate techniques for reducing the dimensionality of data to two or three, at the cost of some error, so as to enable visualization. One such method is *MDS*, or *multidimensional scaling*.

Given a number of points in a, say, twenty-eight-dimensional space, with each group being represented by a vector of twenty-eight morpheme frequencies, MDS arranges the same number of points in a lower dimensional space, *e.g.* in two or three dimensions, with the property that the distances[46] between the points in the lower dimensional space approximate the original inter-point distances.[47] To illustrate MDS with my example of the three morphemes, Figure 11 provides a two dimensional representation, as produced by MDS, of the three-dimensional data.

PCA (Principal Component Analysis) is another method of dimension reduction. Suppose we have a three-dimensional dataset, with each group represented by three morpheme frequencies. Suppose we would like to obtain a

[46] That is, Euclidean distances, which give the length of the straight line connecting two points; see the last two footnotes.

[47] Some MDS algorithms, rather than approximating the inter-point distances, try to preserve the order of the distances, so that greater inter-point distances show up as greater distances in the lower-dimensional representation, without trying to preserve the exact proportions of the original inter-point distances. Incidentally, the original distances may be city-block, Euclidean, or indeed expressed in any other measure of dissimilarity.

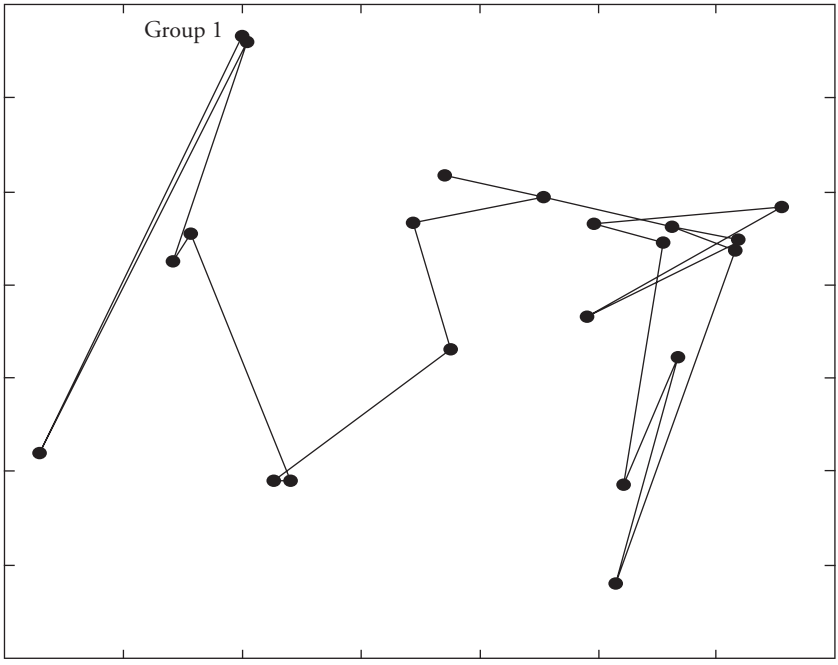

Figure 11. MDS (Multidimensional Scaling). A two-dimensional representation of the twenty-two groups obtained by MDS. The distances between the points in the two-dimensional graph approximate the city-block distances between the points in the original three-dimensional space, depicted above in Figure 10. Group 1 is the point on the upper-left side. Consecutive groups are connected with straight lines. The axes are inconsequential, the only relevant feature being the inter-point distances.

two-dimensional representation of the groups. One way to do so is to take a snapshot with a camera. However, snapshots taken from different angles cover differing amounts of variance in the data. For example, the snapshot in Figure 12 (Left) is much less illuminating in this respect than those shown in Figure 10. The aim is finding a perspective for the snapshot that maximizes the captured variation. This is equivalent to finding a flat surface, say, a sheet of paper, that cuts through the data in an optimal fashion. The two perpendicular edges of such an optimally-placed sheet are the first two principal components. Figure 12 (Right) shows the projection of the twenty-two groups on the plane (flat sheet of paper) spanned by the first two principal components.

In general, the first principal component is defined as the direction along which the data shows the greatest variation. The second principal axis is

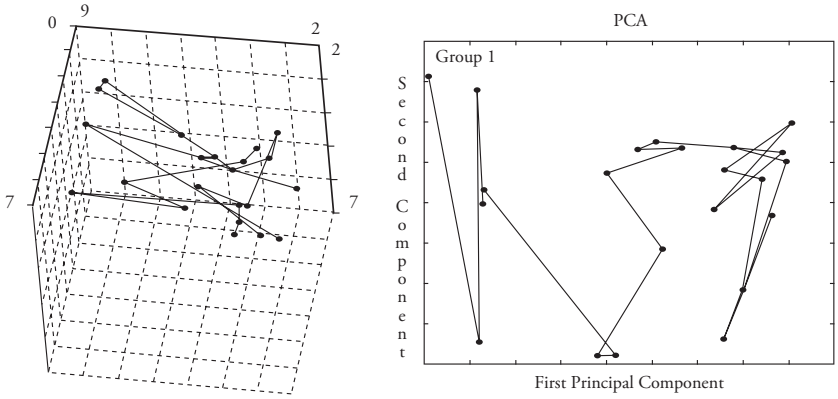

Figure 12. *Left*: Groups observed from an angle that obscures much of the variation in the data. These points are the same as those in Figure 10. *Right*: PCA is tantamount to viewing the dataset shown on the left side from an optimal perspective, *i.e.* one that maximizes the observed variation in the data. The horizontal axis represents the first principal component (*i.e.* the direction of the greatest variation), and the vertical axis the second principal component (*i.e.* the direction of the second greatest amount of variation). These two principal components together account for 91 % of the variance. Variance along the third principal component, that jutting out of the page, is merely 9 % of the total. Group 1 is the point at the upper-left corner. Consecutive groups are connected with straight lines.

chosen from among axes perpendicular ("orthogonal") to the first principal component. Of all such axes, the one is chosen along which the data show the greatest variance. The third principal component (if any) is chosen as perpendicular to the first two, again with the criterion of maximal variance. One may continue in this manner until one has as many principal components as the number of axes (or dimensions) in the original dataset. However, if one desires a lower dimensional representation, one retains fewer principal components than that.

In the example shown in Figure 12 (Right), I have retained two principal components. The third principal component, deliberately omitted, is perpendicular to the two shown, thus jutting out of the page towards us. Therefore, the true position of each point is a bit directly above or below the point on the sheet, and this information is lost. However, the variation along this omitted component accounts for only 9 % of the total variance in the data, the first two components accounting for 91 % of it. In sum, I have successfully reduced

the number of dimensions, though at the cost of failing to capture 9 % of the variance. Such a "loss" is not a bad thing if it eliminates noise in the data that tends to obscure larger trends.

5. Morphemes: Weighting and Weight-Optimization

Basic Concepts: Features, Feature Vectors, Textual Distances

The rest of this essay is devoted to analyzing relative frequencies of morphemes. Henceforth I sometimes use the word "feature" to refer to a morpheme. A morpheme, it will be remembered, is a word or part of a word, delimited by hyphens or spaces in the Wintner-Talmon transliteration of the Qurʾān. For example, the transliterated word *l-raḥmān-i* consists of three morphemes. On average, there are slightly over two morphemes per word.

In this essay I work with three separate lists of morphemes. I now introduce the first two. Feature List A, given in Table 5, consists of the top twenty-eight most common morphemes in the Qurʾān. It can be seen that these features are ones that can be used independently of the subject matter at hand; that is, they are in principle *non-contextual*. They also tend to be *function* features (pronouns, case-endings, etc.), as opposed to *content* features.

Feature List B, provided in Table 6, consists of 114 morphemes that are function features or otherwise are relatively non-contextual, and which are not included among the top twenty-eight most frequent features. These are among the most frequent 536 features in the Qurʾān. I eliminated many elements such as *qāla, arḍ, yawm, qālū*, and *ʿaḍāb* that are firmly tied to specific themes. I did so in a manner that eliminated the possibility of cherry-picking.[48] The idea is to reduce the influence of subject matter in favor of style. In addition, it is important to note that these features occur in the Qurʾān far less frequently than those in Feature List A. Tests indicate that a medium-frequency feature tends to convey less information about style than a high-frequency one. To compensate for this effect, it is important to take a larger number of features, hence the larger size of List B.

[48] The fact that I eliminated some morphemes may raise the suspicion that I picked features that guaranteed the desired result. I eliminated the possibility of such cherry-picking by first removing morphemes on the grounds explained above and then proceeding with the analysis without going back to modify my choices of morphemes. Thus, neither at the moment of elimination nor later did I come to see if or how the results changed. At any rate, the results are too robust too be alterable by the removal or addition of several morphemes. One could randomly add or subtract many morphemes without changing the overall picture.

Table 5. ***Feature List A, consisting of the twenty-eight most frequent morphemes in the Qurʾān. The features are listed here in order of descending total frequency, ranging from 9567 occurrences in the Qurʾān for*** **wa** ***to 1080 for*** **lladḏīna.** ***The third and fourth columns give the relative frequencies of each feature respectively in Groups 2 and 3. For example,*** **wa** ***makes up 6.12 % of the morphemes in Group 2.***

	feature	grp 2	grp 3		feature	grp 2	grp 3		feature	grp 2	grp 3
1	*wa*	6.12	5.23	11	*llāh*	0.27	0.37	21	*īna*	0.68	2.14
2	*i*	5.42	5.04	12	*mā*	1.70	1.58	22	*ka*	1.24	1.04
3	*l*	6.78	6.84	13	*in*	2.21	2.76	23	*hi*	1.01	0.59
4	*u*	4.41	3.93	14	*bi*	1.12	1.37	24	*li*	0.97	0.60
5	*a*	3.33	3.87	15	*un*	1.55	1.85	25	*hā*	1.24	0.91
6	*an*	4.30	1.37	16	*la*	1.24	1.72	26	*ʾinna*	1.05	1.20
7	*min*	0.78	1.48	17	*kum*	0.78	0.67	27	*nā*	0.31	0.73
8	*ūna*	1.36	3.01	18	*hu*	2.01	1.57	28	*alladḏīna*	0.23	0.33
9	*fa*	3.14	2.40	19	*lā*	1.01	1.09				
10	*hum*	0.89	1.54	20	*fī*	1.01	1.02				

Table 6. ***Feature List B, consisting of 114 either relatively non-contextual morphemes or function morphemes, selected from among the top 536 most frequent morphemes in the Qurʾān but excluding those in Feature List A. The features are listed here in order of descending total frequency, ranging from 1050 occurrences for*** **ʾin** ***to 20 for*** **laday.** ***An initial*** **hamza** ***is indicated with an apostrophe.***

ʾin, him, ū, rabb, man, ʿalay, ʿalā, ʾillā, ʾan, ī, ʾa, huwa, ʾilā, ḏālika, ʾiḏā, kāna, ʿan, qad, kull, nī, yā, lam, ʾanna, ʾiḏ, ṯumma, qul, ʾilay, llaḏī, ʾaw, šayʾ, law, kānū, bayn, qabl, hāḏā, ʾayy, ʾūlāʾika, baʿd, kuntum, ʿind, anna, mim, lammā, baʿḍ, maʿa, ʾamr, ġayr, dūn, ḥattā, hunna, n, ʾam, ʾantum, bal, laʿalla, sa, ayni, lan, ā, awna, āni, hal, naḥnu, kayfa, takūn, ʾanta, ni, ya, ḍ, miṯl, ʾaḥad, laysa, yakūn, lākin, ay, aw, ʾanā, llatī, hinna, hiya, iy, lākinna, na, ʾammā, ḏāli, ʾal, ʿam, fal, hāḏihi, unna, hāʾulāʾi, tarā, ʾūl, sawfa, tilka, kunta, kānat, kallā, yakun, ḏā, ḥayṯ, ʿalā, ḏāt, ʾim, rijāl, ʾannā, ki, wal, ʾiyyā, ma, balā, takun, kam, laday.

Given a feature list, any passage or any group of passages may be represented by a list of numbers consisting of the relative frequency counts of each feature in that text. Such a list of numbers is called a feature relative frequency count vector, or feature vector for short. Table 5, for example, provides the feature vectors for Groups 2 and 3 relative to Feature List A. The concept of a relative

frequency count may be illustrated with an example. Take the following verse, in which the morphemes are separated by hyphens: *wa-qurʾān-an faraqnā-hu li-taqraʾ-a-hu ʿalā l-nās-i ʿalā mukṯ-in wa-nazzalnā-hu tanzīl-an* (Kor 17, 106). In this verse there are twenty-one occurrences of morphemes. The feature *wa* makes two appearances; so, its relative frequency count is two divided by twenty-one, or 9.5 %. This number is the relative frequency count of *wa* in this short passage.

The task in the following sections is to use feature vectors to examine the distances between groups, where I use the city-block distance as a measure of the inter-group similarities. This section addresses two issues. First, it considers whether the different features must contribute equally to the calculation of inter-group distances, or whether they should be "weighted" differently. In my discussion of weighting, I will explain the concepts of normalization, standardization, and feature weight optimization. This last involves a new methodological contribution to stylometry. Second, this section tests the utility of Lists A and B by dividing some *sūra*s into halves and checking if the halves display similar frequency count profiles.

Feature Weights: Normalization and Standardization

As one goes down the list of features in Table 5, the frequencies fall sharply. This rapid decline raises a question. Remember from the last section how one calculates the distance between two groups, say, Groups 2 and 3: first, for each feature, one takes the difference in its frequency count. For example, for the first item, *wa*, we get 6.12-5.23 = 0.89. Then one adds up all these differences (for all the features), the sum representing the distance between Groups 2 and 3. Now, as one goes down the list, and as the magnitudes of individual frequency counts decrease, so do the differences. Therefore, items lower in the list contribute significantly less to the sum that defines the distance than do items higher on the list. With feature lists much larger than just twenty-eight items, this discrepancy becomes even more pronounced, with features at the bottom of the list barely contributing anything.

In order to make different features count about the same, one may choose to magnify some of them to make them all have comparable magnitudes before adding them up to get the overall distance. For example, it makes sense to multiply smaller-magnitude features by larger numbers. In general, when one multiplies a feature frequency by a number to adjust its contribution, this number is called a feature *weight*. The list of weights for all the features is called the *weight vector*. One common way of weighting is *normalization*. It means dividing each feature vector by its average magnitude. In the case at

hand, it means dividing a morpheme frequency by the mean value of the frequencies of that morpheme over the twenty-two groups. In this scheme, since smaller-magnitude features are divided by smaller numbers, and larger-magnitude features by larger ones, the feature frequencies become comparable. Another method is *standardization*. It involves dividing the values of a feature by the standard deviation of the feature (over all the groups).[49] Both methods equalize the features. The difference is that normalization happens to give greater weight to features that vary significantly over the groups. Incidentally, in both procedures, it is customary to subtract the feature mean from feature values before multiplying by the weight.

Weight Adjustment

The ultimate goal is to see how similar or different Bazargan's groups are stylistically. The dissimilarity of two groups is defined as the sum of the differences of morpheme frequencies in those two groups. But to simply add up these differences is to give all morphemes the same weight. Is it possible, however, that one morpheme is more important than another and should therefore be given greater weight? *Weight adjustment* refers to the above-mentioned process of multiplying the frequency counts of a feature by a number (coefficient) so as to diminish or boost its contribution to the calculation of distances. If one does not adjust a feature, then the coefficient is simply 1, and the feature vector is kept "raw". As mentioned above, the list of weights for all the features is called the *weight vector*. So, the weight vector for raw features is a list of 1's.

I have shown how weight adjustment can remedy the tapering off effect. However, there are various other reasons why one may want to adjust the weights, and here are three.

First, note that the behavior of a higher-frequency feature can have greater statistical significance than a medium-frequency one. The frequency of a feature that occurs a relatively small number of times in a group is more susceptible to chance oscillations and may thus be statistically less significant and more prone to sampling error. To muffle such noise, it may be fitting to give more weight to the most frequent features.

[49] The standard deviation of a morpheme frequency count is a measure of how much, on average, the frequency of the morpheme differs from its mean frequency. A small value for the standard deviation means that the morpheme frequency tends to stay about the same in the different groups, while a large value indicates greater diversity.

Second, there are cases where two features tend to go together in a text for reasons of linguistics. For example, in the Qur'ān the verb "to be" (morphemes: *kāna, yakūn, yakun*, etc.) correlates strongly with the morphemes representing accusative case endings. A large fraction of the accusatives in the Qur'ān are predicates of the verb "to be"; and conversely, the verb "to be" is typically followed by an accusative. This correlation simply reflects the grammar of the language. So, to give equal weight to "to be" and the accusative case endings would be to count twice what is essentially one grammatical phenomenon, thus possibly hurting precision. One may deal with this by reducing the weights of such elements.

Third, phenomena that are suited to statistical description will typically produce outliers. There will thus be a small number of morphemes whose behavior will be highly exceptional. To improve precision, it would be appropriate to remove these, which is to give them weights of zero.

From this discussion, it must be clear that feature weight adjustment, if based on *a priori* reasoning, can be a highly complex and time-consuming process. Such complexity probably explains why in practice weighting is usually limited to feature normalization or standardization. There is, however, a simple and practical way out that makes weighting automatic. It involves replacing all *a priori* reasoning about weights with *a posteriori* optimization of weights. This technique can achieve higher accuracies than either normalization or standardization.

Weight Optimization

All features are not equally effective discriminators of chronology. Some will be more important than others. We wish to assign each morpheme a weight reflecting its degree of effectiveness. But how can one measure the effectiveness of different features? Rather than speculating in an *a priori* fashion, it is possible to determine this empirically. One can measure the effectiveness of a morpheme by seeing how well it performs at the task of assigning to the same cluster texts that we already know belong together. The basic idea is simple, and can be illustrated with the following hypothetical example: suppose we take a long, coherent, and self-contained *sūra* that we think represents a unified text from one time period. Let us divide it into two halves: say, we divide the *sūra* in the middle—or, say, we combine its odd verses into one text, the "First Half", and combine its even verses into another text, the "Second Half". Now we have two texts that we know belong together. We then pick two morphemes, say *wa* and *fa*, and ask how effective they are as discriminators of style. Looking at the frequencies of *wa* and *fa* in these two texts, suppose one finds that *wa* makes up respectively 6.1 % and 6.2 % of the morphemes in the

two halves, while *fa* forms respectively 2 % and 14 % of the morphemes in them. Note that the percentages of *wa* in the two halves are close to each other, as compared to those of *fa*. One concludes that *wa* is a better indicator of chronology than *fa*, since two passages that date from the same time have similar frequencies of *wa* but not similar frequencies of *fa*. Therefore, from now on, when one computes the distances between passages by summing up the differences of morpheme frequencies, one assigns more significance to the differences in the frequency of *wa*. One does so by multiplying that difference by a relatively large number before adding it to other morpheme differences. So, different morphemes are weighted differently as their contributions are added up to obtain the overall stylistic distance between two passages.

If one takes a number of unified texts, and divides each text into two halves, then one may measure the success of different weighting schemes in indicating the affinity of the pairs of halves. A weighting scheme is more successful if it does a better job of clustering together halves that truly belong together. In fact, one may try many different weighting schemes and pick the most successful one. This gives rise to the following optimization problem for finding the best weighting scheme: find a weight vector that achieves the best clustering success. To perform this optimization, one needs two things: a set of training texts and a measure of clustering success.

Training texts (see Table 7) are bisected unified texts used in the process of optimizing feature weights. That is, the weights are chosen for superior performance specifically on the training texts. It is important that the training texts be representative of the whole corpus; otherwise good performance on the training texts may not lead to good performance in general. Normally, the more numerous the training texts are, the more representative they will be of the larger corpus. Moreover, medium and large texts usually yield better optimized weights than small ones. To compile the eighteen training texts used here, I started with all twelve intact *sūra*s of over 570 words, then added five large contiguous passages, each unified in Bazargan's reckoning and each forming part of a *sūra*.[50] Finally, I combined several smaller passages into one text (text number 17), having determined from old-fashioned stylistic analysis that they probably belong together.[51]

---

[50] These five texts I chose from six that I had picked randomly, though with a view to choosing large passages. Of the six texts, I removed the one that displayed the largest differences in the styles of its two halves. This *relative* imbalance could be a sign of one of two things. It could be that the halves belong to different times, or it could be that they are contemporaneous but unusual in their difference. In either case, the exclusion is not problematic.

[51] I examined the occurrences of unusual phrases, words, and themes to arrive at the hypothesis that *sūra*s 48, 58, 63, 66, and *sūra* 9, verses 71-96 (respectively Blocks 170, 181, 186, 152, and part of 169) are close in time. If I am wrong about this, the error will not be fatal, since each

Table 7. ***Training Texts. One can test a method by seeing how well it joins together texts that belong together. To that end, eighteen texts are divided into halves and used alternately for optimizing ("training") the weight vector and for testing it. Texts 1, 3, 5, 6, 8, 9, 10, 11, 12, 13, 15, and 16 are complete* sūras*—namely,* sūras *37, 36, 21, 25, 27, 11, 29, 12, 42, 8, 57, 13. Texts 2, 4, 7, 14, and 18 are parts of* sūras. *Text 17 is a composite of several probably relatively contemporaneous passages, namely all of* sūras *48, 58, 63, and 66, and part of* sūra *9.***

Text no.	Halves definition (*sūra*) block: verses	words
1 (a)	(37) 58: 1-87.	432
	(37) 58: 88-182.	433
2 (b)	(26) 65: 52-139.	399
	(26) 65: 140-227.	417
3 (c)	(36) 88: 1-43.	366
	(36) 88: 44-83.	363
4 (d)	(23) 90: 12-62.	502
	(23) 90: 63-118.	493
5 (e)	(21) 94: 1-55.	591
	(21) 94: 56-112.	583
6 (f)	(25) 102: 1-37.	452
	(25) 102: 38-77.	444
7 (g)	(20) 103: 53-94.	493
	(20) 103: 95-135.	475
8 (h)	(27) 114: 1-45.	581
	(27) 114: 46-93.	577
9 (i)	(11) 118: 1-57.	969
	(11) 118: 58-123.	976
10 (j)	(29) 126: 1-35.	497
	(29) 126: 36-69.	480
11 (k)	(12) 130: 1-55.	890
	(12) 130: 56-111.	903
12 (l)	(42) 138: 1-23.	417
	(42) 138: 24-53.	444
13 (m)	(08) 143: 1-41.	628
	(08) 143: 42-75.	615
14 (n)	(10) 160: 1-30.	609
	(10) 160: 31-70.	624
15 (o)	(57) 161: 1-16.	292
	(57) 161: 17-29.	282
16 (p)	(13) 168: 1-20.	422
	(13) 168: 21-43.	431
17(q)	(63) 152: 1- 5. (9) 159: 38-54. (9) part of 169: 71-84. (48) 170: 1-17. (58) 181: 1-9. (66) 186: 1-6.	1305
	(63) 152: 6-11. (9) 159: 55-70. (9) part of 169: 85-96. (48) 170: 18-29. (58) 181: 10-22. (66) 186: 7-12.	1263
18 (r)	(09) 177: 1-20.	361
	(09) 177: 21-37.	371

half-text contains half of each of these passages, so that the two half-texts should be assigned together anyway even if the individual passages belong to different times. After I had inferred the contemporaneity of these passages, I consulted the chronological sequence of *sūra*s ascribed to

As a measure of how well a given weighting scheme assigns together the two halves of a text, I introduce the quantity called *Individual Clustering Quality* (ICQ). Given the two halves of a bisected text among a collection of such texts, the ICQ of a weighting scheme with respect to this text is defined thus:

$$\text{Individual Clustering Quality} = \frac{\text{Distance of the text from the nearest } m \text{ neighbors}}{\text{The distance between the two halves of the test}}$$

where normally one chooses $m = 1$ or $m = 2$, representing one or two neighbors. The idea behind the denominator, which one may call the "*intra*-text distance", is simple enough: a scheme that places the two halves closer to each other, thus resulting in a smaller denominator, has a higher ICQ. (As for how to calculate this distance, recall that each half is represented by a weighted feature vector. Thus, the distance between two halves is simply the city-block distance between their respective weighted feature vectors.)

On the other hand, the numerator, which one may call the "*inter*-text distance", reflects how well the clustering procedure distinguishes this text from other texts. It measures the "distance" of the text at hand (both halves of it) from the nearest text(s) (represented by their respective halves). The greater the distance from the nearest texts is, the higher the success of the clustering method in distinguishing this pair from the other pairs. This concept turns on the notion of the "distance" between two pairs of halves, which I define as the distance between their respective midpoints, where the midpoint of a pair of halves is obtained by averaging their feature vectors. For example, if the frequency of *wa* is 5 % in the first half and 6 % in the second half, the midpoint vector will have 5.5 % for its frequency for *wa*. Such averaging is used also to obtain all the other feature frequencies, thus defining a full feature vector representing the midpoint. With the midpoint feature vectors in hand, one may obtain the distances between the midpoint of the text at hand and the *m* nearest midpoints (*i.e.* those midpoints with the smallest distances from the midpoint at hand). The average of these distances forms the numerator. To help with the comprehension of these ideas, Figure 13 illustrates these definitions graphically.

Having defined the notion of Clustering Quality for *an individual* text, one may now extend the concept to *a collection* of training texts. I calculate the

---

Ǧaʿfar al-Ṣādiq, and noted that they are also located near each other within this fragment of his sequence: **63**, **58**, 49, **66**, 61, 62, 64, **48**, **9**. (For an analysis of the al-Ṣādiq sequence and those ascribed via different *isnād*s to Ibn ʿAbbās, see Bāzargān, *Sayr*, vol. II, p. 192-203 / 557-69. For a discussion of the sources in which these sequences and some others are found, see Rāmyār, *Tārīḫ-i Qurʾān*, p. 661-7.)

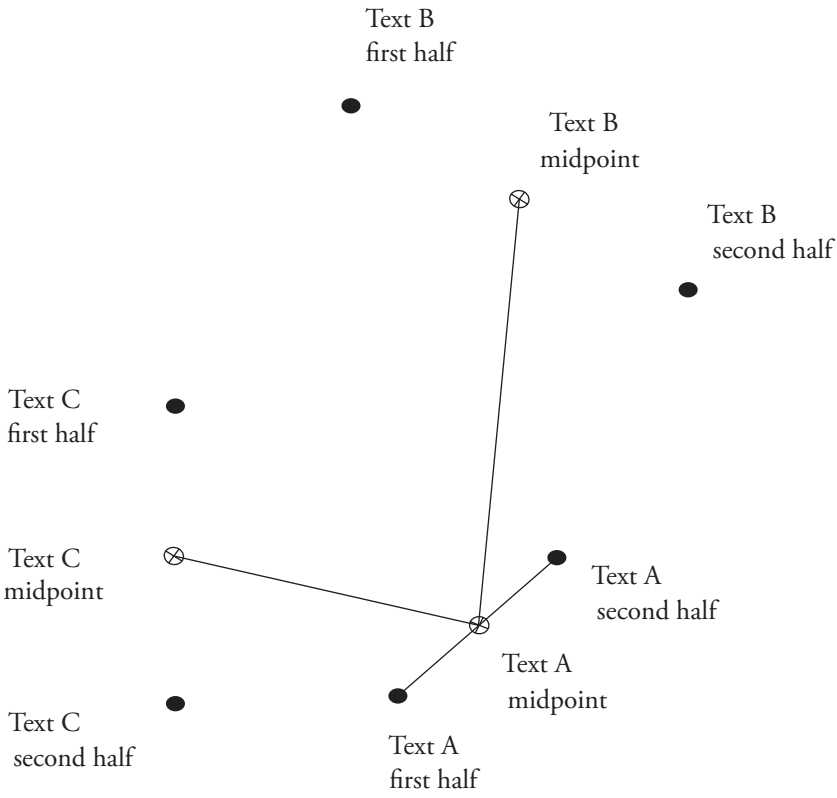

Figure 13. Individual Clustering Quality (ICQ) is a measure of how well the two halves of text A cluster together. The closer the two halves of Text A are together, and the farther apart Text A is from other texts, the higher the ICQ for text A. How does one calculate the ICQ for Text A? Suppose the two nearest neighbors are texts B and C as shown. The denominator of ICQ is the distance between the two halves of Text A, represented here by the length of the line segment connecting them. The numerator depends on the choice of the number of neighbors to consider, m. If $m = 1$, then the numerator is the distance from the midpoint of Text A to the midpoint of the nearest text, namely Text C. This distance is represented by the line segment connecting the two midpoints. If $m = 2$, then it is the average of two distances, namely (1) the distance from the midpoint of text A to the midpoint of text B, and (2) the distance from the midpoint of text A to the midpoint of text C. These two distances are represented by the line segments connecting the midpoint of Text A to those of Text B and Text C.

inverse of the ICQ for each member of the set, take the average of these values, and invert this average to obtain the *Total Clustering Quality* (TCQ). This is basically a form of averaging, but one involving taking inverses before and after.

One may now pose the optimization problem as follows: *find a vector of weights, i.e. a list of coefficients for the features, such that it results in the largest Total Clustering Quality attainable*. This is tantamount to finding the weights that do the "best job" of putting together training texts that belong together and setting apart those that do not. "Best job", of course, requires a measure. To that end, I defined the TCQ. A higher value of TCQ means that pairs of half-texts are closer to each other and father apart from other texts. "Weight optimization", therefore, means finding weights that maximize the TCQ. This optimization can be performed within a software environment for technical computing such as Matlab. A solution to this problem is called an *optimal*, or *optimized*, *weight vector*. The optimal weight vectors for Lists A and B are given in Table 8.

Table 8. ***Weights optimized with m=1 for feature lists A and B.***

The optimal weight vector for **Feature List A**: (91, 0.080, 0.18, 6.5, 0.24, 87, 0.03, 0.92, 50, 0.22, 35, 23, 0.15, 169, 85, 317, 0.02, 0.01, 0.03, 0.06, 181, 57, 90, 0.27, 0.06, 209, 83, 2.8)
The optimal weight vector for **Feature List B**: (13.9, 101, 0.046, 5.7, 0.29, 0.25, 0.048, 0.41, 0.05, 0.14, 0.46, 0.23, 0.59, 0.25, 0.96, 148, 0.89, 140, 1.23, 0.48, 94, 2.3, 3.5, 4.6, 0.1, 30.5, 0.1, 30.6, 0.7, 1.1, 0.6, 208, 0.4, 1.2, 38, 1.5, 0.2, 1.2, 0.6, 371, 281, 228, 4.6, 0.1, 16.3, 0.3, 0.3, 0.3, 1.3, 3.5, 532, 9.7, 599, 1.0, 0.9, 1.7, 1.1, 570, 60, 2.4, 1.1, 1.6, 741, 504, 166, 8.2, 0.4, 0.3, 3, 21, 272, 197, 247, 15, 81, 81, 126, 66, 102, 20, 1.7, 57, 214, 0.7, 0.9, 350, 1.7, 128, 1.3, 312, 99, 322, 7.6, 204, 391, 146, 80, 62, 0.5, 723, 50, 850, 109, 0.7, 0.7, 3.4, 1.0, 5.9, 13.0, 1.0, 8.3, 15.4, 0.5, 570)

How Well are the Halves Clustered?

The dendrograms in Figure 14 depict the *agglomerative hierarchical clustering* of the eighteen pairs of half-texts with normalized features from List A (top) and with standardized features (middle). Each text is designated with a letter of the alphabet. The closest pairs of half-texts are clustered together in the first round, forming the lowest-level clusters. Next, the closest pairs of objects (half-texts or clusters) are grouped together in the second round, and so on,

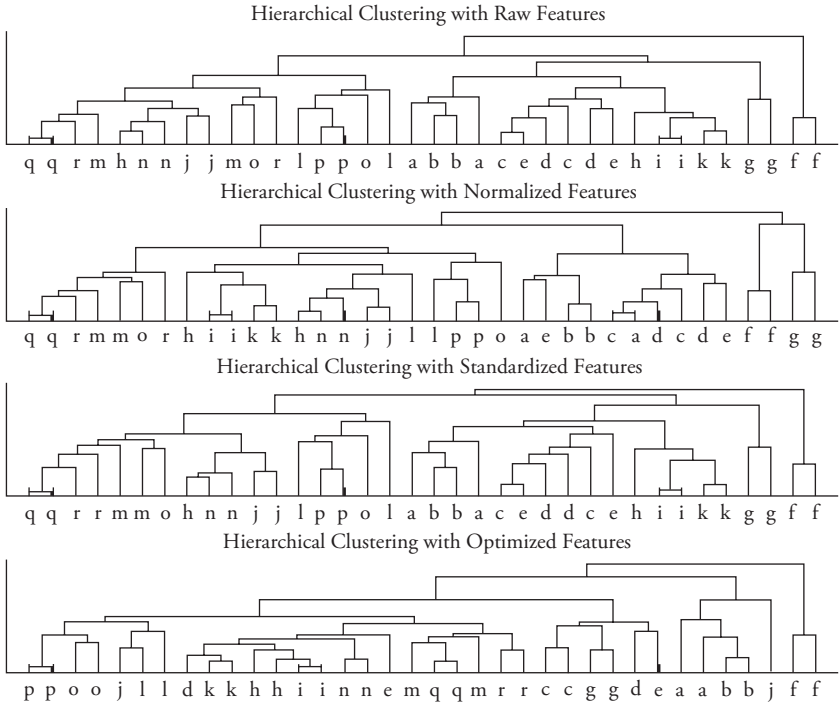

Figure 14. Hierarchical clustering with Feature List A, with raw, normalized, standardized, and weight-optimized (*m*=1) features. The halves of Text 1 are designated with the letter "*a*", the halves of Text 2 with the letter "*b*", and so on. The length of a U-shaped line is a measure of distance. Raw features yield eight exact matches and two near matches (*a, n*), normalized features yield nine exact matches (*i.e.* with two halves in the same lowest-level cluster) and one near match (namely, Text *n*), while standardized features yield eight exact matches and four near ones (*a, d, n, r*). Where there is no match, generally pairs of half-texts that belong to each other are still part of the same mid-level clusters. Thus, texts that Bazargan considers as close to one another indeed tend to be nearer one another. Optimized weights are designed to maximize the number of matches, and they do, yielding twelve exact matches and one near match (Text *a*).

until everything is subsumed in a single super-cluster. The lengths of the U-shaped lines indicate the distances between the objects (texts or clusters).[52] With standardized features, we observe twelve exact or near matches in twelve out of the eighteen pairs of half-texts. An exact match is when the two halves that truly belong together are clustered at the lowest level, in the first round. Even where there is not a match, generally two half-texts that belong together inhabit the same mid-level cluster. As stylometric studies go, such level of performance is outstanding, especially considering the small sizes of the texts, the crowded field of choices (with thirty-two items), the fact that the texts are all from the same work, and the fact that some of the texts may actually be contemporaneous.

In addition, it is impressive that the texts that Bazargan considers near in time tend to be placed nearer to each other. Thus *a* is closest to *b*, and together they are closest to the cluster {*cde*}. For later texts, there is a similar tendency, but less exactly and only in a broad way. If one uses brackets to indicate the nearness of texts, the overall clustering scheme looks something like this:

f, *g*, {{*ab*}{*cde*}}, { {{*ik*} *h* {*jn*}} {*lop*} } {*mqr*}

The bottom dendrogram in Figure 14 depicts the clustering of the half-texts when represented by weight-optimized feature vectors. The key point here is that there are several more matches than in the previous weighting schemes. Such improvement is expected. The clustering scheme is broadly compatible with the previous cases, the most interesting difference perhaps being that now Text *g* is firmly joined to *cde*.

Figure 15 depicts the dendrograms when the half-texts are represented with feature vectors from List B. In the raw, normalized, and standardized cases, the quality of clustering seems inferior to the results of List A (see the top three dendrograms). However, weight optimization helps yield performance that is nearly as good as before (bottom dendrogram).

Despite its visual appeal, hierarchical clustering is not an infallible way of judging clustering success. In some situations, texts that are close to each other may end up in clusters that are far apart.[53] As a more reliable guide, I have defined a *clustering score*. Table 9 provides the scores. As expected, optimized weights perform much better for both Lists A and B.

---

[52] The distance between two half-texts is defined as usual. The distance between two clusters is defined here as the average distance between all pairs of half-texts in them.

[53] For example, suppose A is the closest text to B, but C is closer than B to A. Then A and C will form the cluster {A, C}. In the next step, likewise, {A, C} may be joined with another item rather than B, and so on.

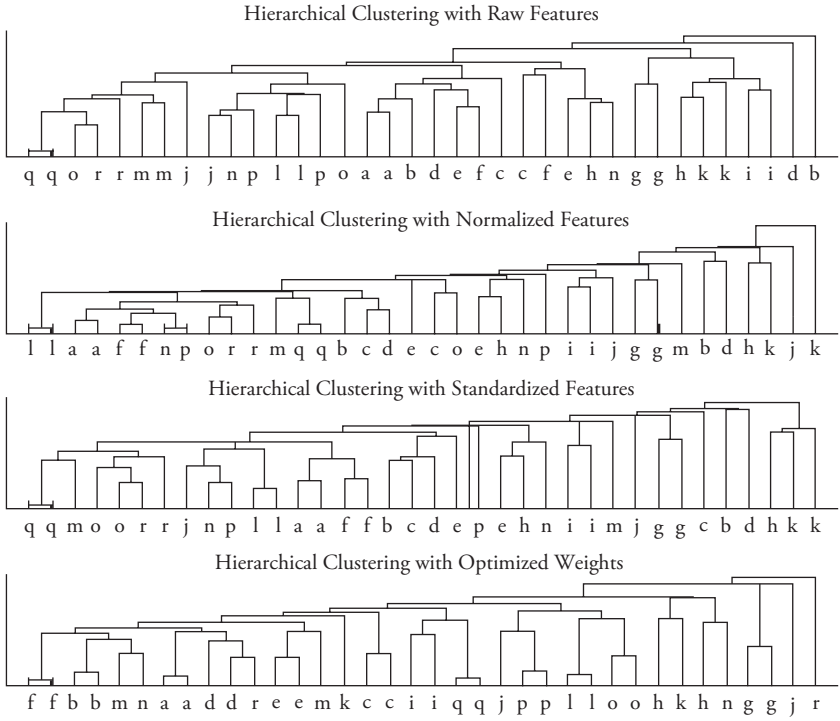

Figure 15. Hierarchical clustering with Feature List B, with raw, normalized, standardized, and weight-optimized (*m*=1) features. Raw features yield six exact matches and two near matches (*k, r*), normalized features six exact matches and one near match (*r*), standardized features six exact matches and three near ones (*k, o, r*), and optimized weights eleven exact matches and one near match (*d*).

Table 9. ***Clustering Scores. The scores below are obtained as follows. First, for each half-text one records a score indicating the proximity of its twin half. The score is 1 if the twin half is the nearest text, ½ if it is the second nearest text, ⅓ if it's the third, and so on. The scores for all half-texts are added up, and the total is expressed as a percentage of the maximum possible score (which would be 36).***

Weighting	Feature List A	Feature List B
Raw	64.9	53.4
Normalized	67.3	44.8
Standardized	64.8	56.0
Optimized	80.9	77.2

Optimized weights are contrived to yield superior clustering for training texts. Indeed, as seen above, they fulfill this expectation. But do they give superior results when applied to texts not used in training? One cannot know that without testing them. That test is called cross-validation.

Cross-Validation as a Test of Weight Optimization

Is there any reason for believing that weight optimization performs better than other weighting schemes—better than if one used raw, normalized, or standardized features? Weight optimization is a form of learning: learning how to weight the different features. But it is learning from a limited set of texts, namely the training texts. We just saw that optimized weights perform better than other weighting schemes at clustering the *training texts*. But this is hardly a surprise. Optimized weights perform well on these training texts because that is precisely what they are designed to do. The real question is whether optimized weights outperform other schemes on texts on which they have *not* been trained. Thus the true test of the merit of optimization is to see how it clusters texts not involved in training. Think of an apprentice who has been trained in the mechanics' school to fix three specific training cars. One would like to know if the learning has made him/her better at fixing cars in general. One finds this out by testing the mechanic with a car he/she has not trained on.

Now think of the eighteen texts as cars. You hide the first car, train an apprentice on all the other cars, and then test him on the first car. You then hide the second car, train an apprentice on the seventeen other cars, and then test her on the second car. And so on. Thus, one excludes the first text from training and uses all the *other* texts combined to optimize the weights, *i.e.* by choosing the weights that do the best job of clustering them. One then uses the resulting optimized weights to test how well they join the halves of the first text: one calculates the clustering quality, ICQ, to see if it is higher with optimized weights than with raw, normalized, and standardized features. If ICQ is higher for the optimized weights, that means better performance than the other weighting schemes in this particular case. But maybe this text is an aberration? So, next, one excludes the second text from training instead, trains the weights on all the other texts, and sees how they perform at joining the halves of the second text compared to raw, normalized, and standardized features. One repeats this procedure for each of the texts.

Figure 16 shows the results for Feature List A. Raw, normalized, and standardized features differ among themselves in their performance, but these differences are over-shadowed by the superior performance of optimized weights. Figure 17 shows a similar improvement in the case of Feature List B. Cross-validation, therefore, confirms that optimizing weights improves the

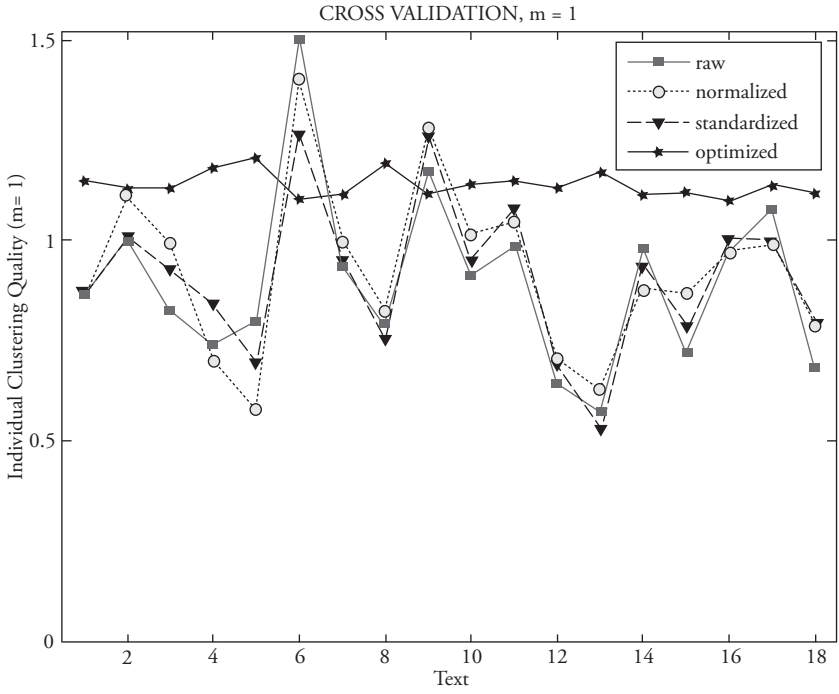

Figure 16. Cross-validation of optimized weights for Feature List A. The plot shows the value of ICQ (*m*=1) for each text for raw, normalized, standardized, and weight-optimized feature vectors. A higher value of ICQ means better clustering. The weights are optimized (with *m*=1) with the text at hand excluded from the training set. The mean value of ICQ is 1.14 for optimized weights (with standard deviation of 0.03). It is significantly lower for the other weighing schemes, namely 0.90, 0.92, and 0.91 respectively for raw, normalized, and standardized features (with standard deviations respectively 0.22, 0.21, and 0.19).

accuracy of multivariate analysis. It can also be seen, however, that optimization does not *always* perform better. The improvement is seen somewhat more consistently in the right half of the graph, in texts 11 and above.[54]

I am now in a position to address a nagging question. Could it be that some of the texts that I took to be unified actually are not, and if so, does that

[54] As will become evident later, style varies more dramatically in groups 1-11. This means that more training texts from this phase may be needed than those included at present, hence the less reliable results. The problem is particularly acute for the first few groups.

invalidate optimized weights? It is not entirely impossible that in a case or two the two halves of a text date from different periods. This, however, is unlikely to have been the case for more than a couple of texts, if that. Otherwise, cross-validation would not have yielded such favorable results for optimized weights, nor would have the dendrograms exhibited clustering of such high quality. The question also arises as to whether halves that were not assigned together should be assumed to be from different times. The answer is no. In other stylometric studies where a collection of texts have been considered simultaneously, even the most effective techniques have not succeeded in clustering texts without error. If two texts are not assigned to the same cluster, that does not necessarily mean that they do not belong together.

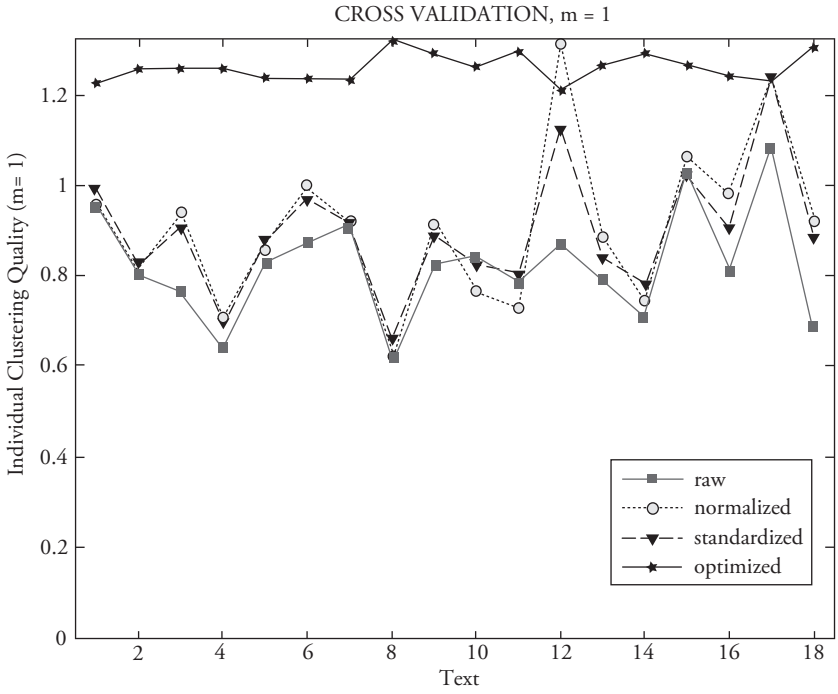

Figure 17. Cross-validation with Feature List B. The weights are optimized (with m=1) with the text at hand excluded from the training set. A higher value of ICQ means better clustering. The mean value of ICQ (m=1) is 1.27 for optimized weights (with standard deviation of 0.03), while mean ICQ (m=1) is 0.85, 0.93, and 0.92 respectively for raw, normalized, and standardized features (with standard deviations respectively 0.13, 0.18, and 0.15).

Section Summary and Conclusion

One may consider different weighting schemes. The weight of a feature is the number by which one multiplies its frequency count to boost or reduce its contribution to the calculation of distances among texts. One particular weighting scheme consists of leaving the frequency counts untouched, or "raw". (This is equivalent to having a weight vector consisting purely of 1's.) Two common alternatives are normalization and standardization. Finally, using weight optimization one may attempt to find weights that most accurately cluster half-texts. One may do so by finding weights that maximize Total Clustering Quality (TCQ). Once the weights are obtained in this manner, testing them using cross-validation verifies them as competitive with other weighting schemes.

This section has given a glimpse of the promise of multivariate techniques as applied to frequency counts of the morphemes of the Qurʾān. Particularly interesting are the results obtained from the list of the top twenty-eight morphemes. I find that for passages that are larger than a certain size—the threshold lying somewhere between 300 and 400 words—these methods, when used with normalized or standardized features, constitute a reliable means of judging the stylistic relationships of texts—reliable, but not flawless. With optimized weights, one might obtain reliable performance for somewhat smaller texts.

The above investigation has implications for theories about the composition of the Qurʾān. The traditional understanding, as embodied in Bazargan's work, acknowledges that the *sūra*s may contain passages from different periods, but it also tends to assume the chronological unity of many *sūra*s and many pericopes. On the other hand, the scholar of the Qurʾān, Richard Bell, took a very different approach. Even in the case of the twelfth *sūra* (Yūsuf), which is normally regarded as unified and coherent, he viewed the *sūra* as a hodgepodge, a patchwork of small fragments belonging to several periods, and the outcome of an extensive process of collection, revision, and interpolation by the Prophet.[55] On Bell's approach, one would deny medium and long passages stylistic distinctiveness or temporal unity.

Richard Bell's general vision does not fit the findings in this section. What has been demonstrated is the stylistic distinctiveness and coherence of passages to a surprising degree—that is, to a degree greater than what stylometric techniques have demonstrated in some other cases in which the questions of

[55] Richard Bell, *A Commentary on the Qurʾān*, eds Bosworth and Richardson, Manchester, University of Manchester, 1991, vol. I, p. 375-406.

authorship and chronology are not in doubt, and where the texts are much longer and the analysis therefore less prone to error. (See especially Figure 14.) The traditional understanding is correct.

The primary aim of this section has been to test and hone the techniques to be used in the rest of the essay. As an incidental bonus, however, a phenomenon has emerged that already lends some support to Bazargan's chronology. One finds that, speaking rather broadly, the dendrogram in Figure 14 places near each other the passages that are close in time according to Bazargan's chronology.

6. Multivariate Analysis (List A): Top Twenty-Eight Morphemes

We now come to the heart of the argument. I showed how several univariate markers of style behave over Bazargan's sequence of twenty-two groups of passages. In particular, mean verse length varies in a smooth fashion. I now compare the groups with the aid of the multivariate techniques described in the last two sections and using Feature List A. This list contains the twenty-eight most common morphemes in the Qur'ān as presented in Table 5 on page 253. Thus, each group is represented by twenty-eight numbers, namely the frequency counts of the morphemes of List A. The goal is to see whether style, as represented by the relative frequency counts of these twenty-eight morphemes, varies smoothly over Bazargan's sequence of twenty-two groups. In other words, do groups that are consecutive or nearby in Bazargan's chronology have similar stylistic profiles? To the extent that they do as judged by different markers of style, there is concurrent smoothness; and any part of the sequence that displays concurrent smoothness is confirmed.

Stylistic dissimilarity is represented using the graphical techniques of MDS and PCA. Each of Bazargan's groups is represented by a dot. The distance between two dots is a measure of the stylistic dissimilarity of the corresponding groups. Thus, two groups that are stylistically similar, in the sense of using the top twenty-eight morphemes with similar frequencies, are placed near each other. The farther two groups are on the diagrams, the more different their stylistic profiles are as measured by how frequently they use the morphemes in List A. As the main question is the presence of smoothness, the aim is to check for relative proximity of consecutive groups.

Figure 18 presents the PCA plot of the twenty-two groups as represented with normalized features from List A. This two-dimensional representation accounts for 64 % of the variance in the data. It gives a broad idea of the true shape of the data. Figure 19 presents graphs obtained using four different methods of MDS.

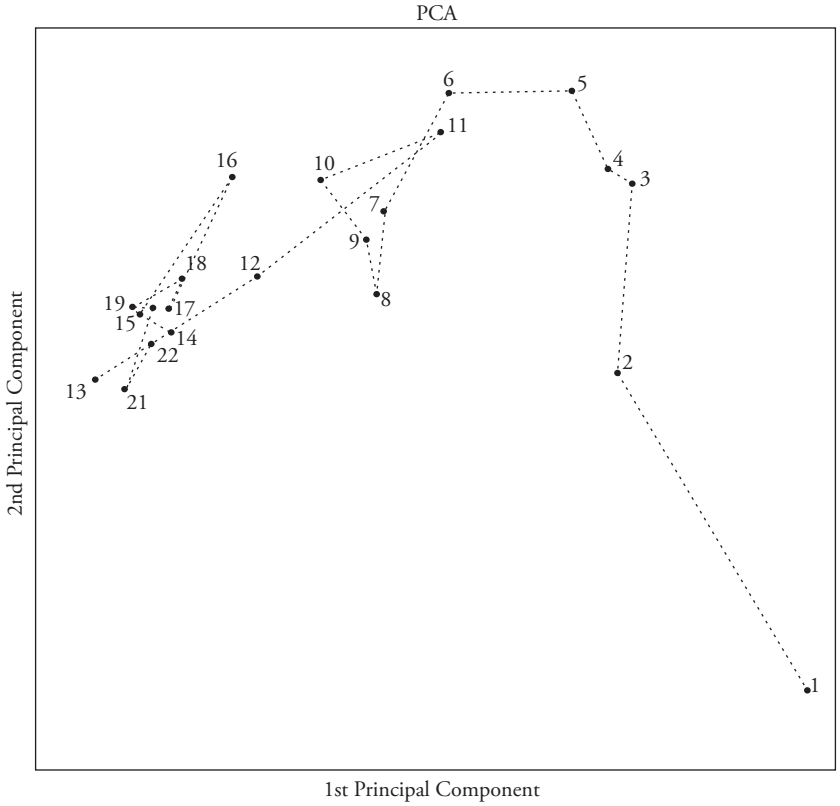

Figure 18. PCA plot of Bazargan's groups as represented by normalized features of List A. Consecutive groups are connected with dotted lines. The two principal components account for 64 % of variance in the data.

The MDS method named "Sammon mapping" yields a "stress value" of 5.4 %. Stress is a measure of how well the distances in the two-dimensional representation match those in the original space of dimension twenty-eight. Stress values of 2.5 %, 5 %, and 10 % would be considered respectively excellent, good, and fair reproductions of the distances.[56] Thus, Sammon mapping gives a close approximation of the distances. The PCA and MDS plots obtained using standardized weights are shown in Figure 20. Table 10 in the appendix provides the actual distances among the twenty-two groups (normalized case). Table 12 in the appendix helps with the interpretation of the distance matrices.

[56] B.S. Everitt and G. Dunn, *Advanced Methods of Data Exploration and Modeling*, Exeter, New Hampshire, Heinemann Educational Books, 1983, p. 65.

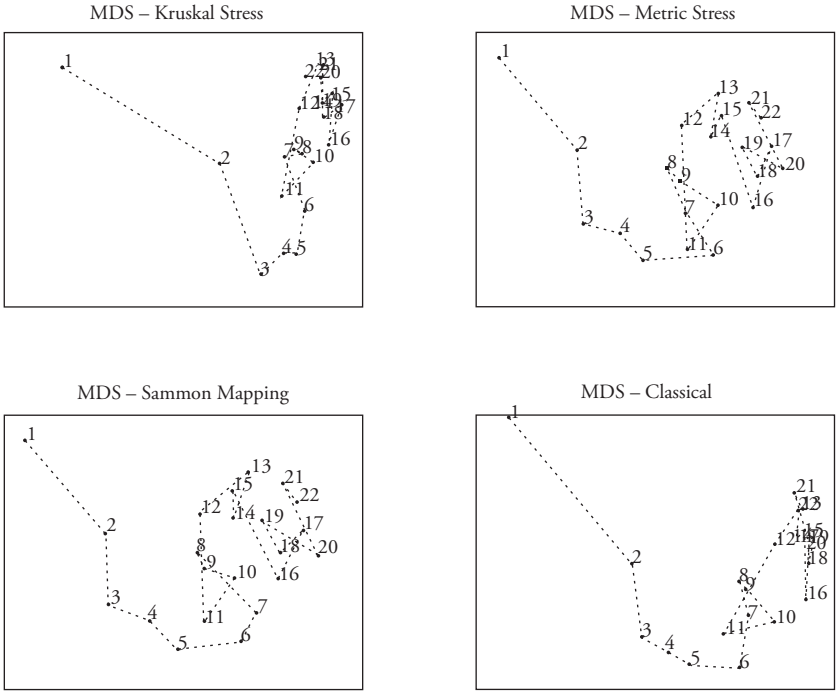

Figure 19. MDS graphs of the twenty-two groups as represented by normalized features of List A. Four different methods of MDS are used. For Sammon mapping, the stress value is 5.4 %, indicating good reproduction of the original interpoint distances, while the others have higher (worse) stress values.

One may now examine the graphs for smoothness. The question is: *are groups that are near each other in Bazargan's sequence also near each other in the graphs?* This is another way of asking: do groups that have similar verse lengths use the most common morphemes with similar frequencies? In a broad way, that appears to be the case. One observes a progression in three regimes: (I) Groups 1-6, (II) Groups 7-11, (III) Groups 12-22. (Regime III is the Medinan period in Bazargan's reckoning.) In Bazargan's sequence, as well, these regimes appear in order.

Within Regime I, one observes not only great smoothness, but also a clear progression. Within the other regimes, there is no clear progression, and only a limited tendency for consecutive groups to be near each other. Thus, we have the following clusters of consecutive groups that lie near one another: {8, 9, 10}, {12, 13, 14, 15}, and {16,17, 18, 19}. In the normalized case, Groups 12 and 16 occupy intermediate positions between regimes II and III. In the case of Group 12, there is nothing odd about this. The location of Group 16, however, could be considered aberrant. Group 11 appears a bit out of place as well.

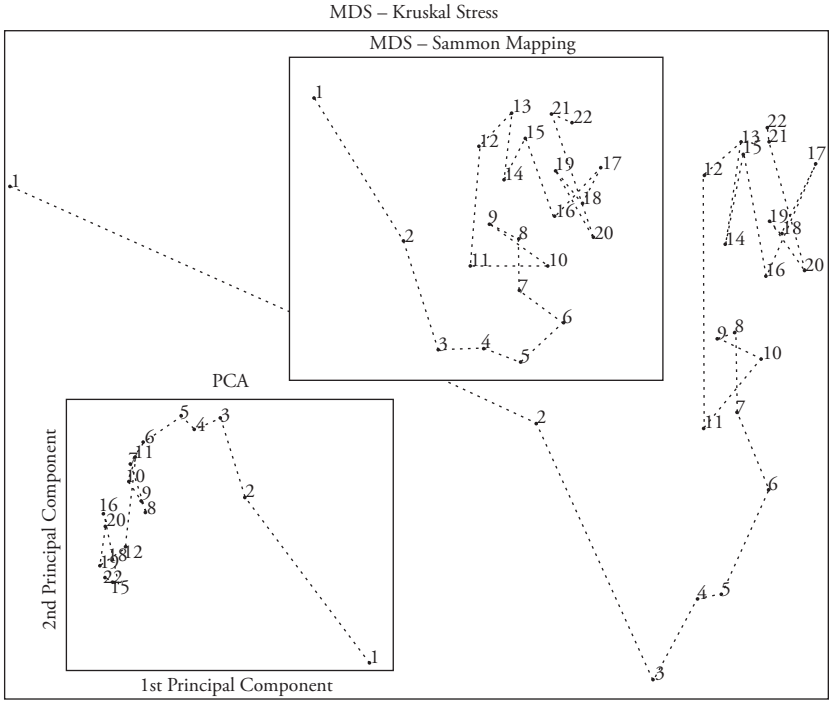

Figure 20. PCA and MDS, List A, Standardized. The first two components in PCA (inset, bottom) explain only 56 % of the variance. MDS yields stress values of 5.8 % (good) and 8.7 % (fair) respectively for Sammon and Kruskal.

Figure 21 provides PCA and MDS graphs of the groups as represented by weight-optimized features. PCA accounts for 67 % of the variance in the data, while MDS yields a good reproduction of inter-point distances (stress=4.7 %). Table 11 in the appendix provides the pairwise distances among the twenty-two groups, and Table 12 in the appendix helps with the interpretation of the distance data.

The first two groups, not shown in the plot, are far away from where one would expect them to appear. This is due to the inadequacy of optimized weights when it comes to the first two groups. This inaccuracy results from the fact that the training set lacks any texts from the first two groups. Now, if the first two groups had been stylistically close to the others, this absence might not have been detrimental. But as the normalized and standardized cases demonstrate, they stand far apart from the rest. This separation means that the

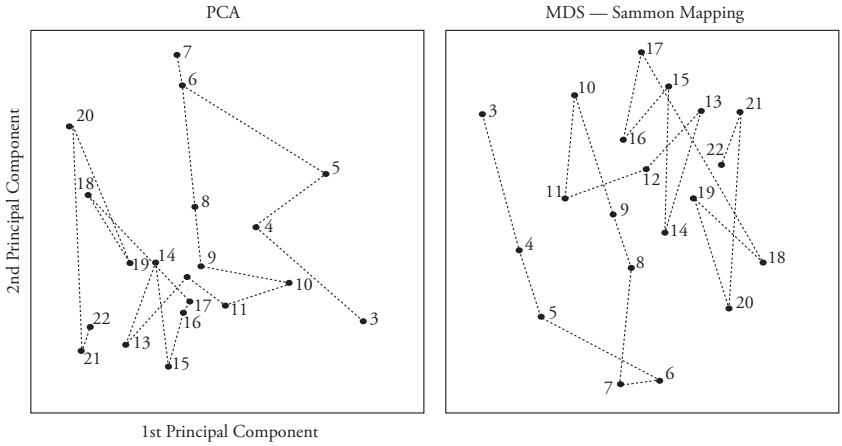

Figure 21. PCA and MDS (Sammon Mapping) representations of Groups 3-22 with optimized weights from List A. The PCA plot explains only 67 % of the variance, while MDS provides a good representation (stress=4.6 %).

optimization process will miss the importance of the morphemes that are distinctive of the first two groups, giving such features weights that are zero or close to zero. So, the first two groups end up being judged on the basis of morphemes that are suited specifically to the other groups.[57] To put all this in less technical terms, for accurate results the training texts must be representative; but in fact they are not representative of the first two groups.

Optimized weights yield a somewhat different picture. The three regimes remain valid. However, the boundary of the first two regimes is now blurred as Group 7 joins Group 6. More interestingly, one now observes hints of smoothness internally in what I called above Regimes II and III. Certain consecutive groups happen to be neighbors: {9, 10, 11}, {12, 13, 14}, {15, 16, 17}, {18, 19, 20}, and {21, 22}. This does not mean that there is either strict smoothness or a steady progression within regime III: note the jump from Group 14 to 15 and especially that from 17 to 18. Unlike in the normalized case, Group 16 does not occupy an aberrant position.

The upshot is this: Consider the following sequence of eight phases: {1} {2} {3} {4} {5} {6} {7-11} {12-22}. Style, as represented by the frequencies of morphemes in List A, varies in a smooth fashion over this sequence of phases. However, it does not vary smoothly within these phases. What this means for

[57] The groups 21-2 are not represented in the training set. Therefore, the results in their cases cannot be fully trusted either. However, they are stylistically close to groups that are well represented in the training set. Therefore, the results can be expected to be not too far off.

concurrent smoothness, and hence chronology, will be discussed after two other multivariate markers of style are considered.

7. Multivariate Analysis (List B): 114 Frequent Morphemes

This section is based on the 114 features that make up List B, which is provided above in Table 6 on page 253. Since this list does not overlap with List A, it can be used for an independent assessment of style. Admittedly, there may be a few features in List B that correlate strongly with some from List A for linguistic reasons, but the effects of such correlations on the final results are negligible, and the two lists can be considered practically independent. Indeed, the overall results in this and the last section are not sensitive to the inclusion or deletion of a few features. Thus if one were to eliminate the most highly correlated features, there would be no important change in the overall results.

Each group is now represented as a vector of 114 relative frequency counts, namely those of the features in List B. Just as was done above for List A, the aim is to check for smoothness over Bazargan's sequence of twenty-two groups.

I have included PCA and MDS graphs for normalized, standardized, and weight-optimized features. Let us examine the graphs. In the standardized (Figure 22, Figure 23) and normalized cases (Figure 24, Figure 25), the three regimes from the last section reappear more or less. The key change is that now the last regime (Groups 12-22) is divided into two parts: {12-19}, {20-22}. These two clusters line up to create a progression. Thus, the groups that for Bazargan start and end the revelation form precisely the extremities of the MDS and PCA representations, with the clusters between them lining up broadly in the expected way. Group 16 occupies an idiosyncratic position that is already familiar from List A. In most graphs, Group 12 occupies an intermediate position between two clusters.

In the graphs of the weight-optimized case (Figure 26), Groups 1 and 2 must be ignored as before. One of the Groups 3, 4, or 5 has an aberrant position, but it is difficult to tell which. Thereafter, there is a great deal of smoothness, with consecutive groups near each other. Groups 16 and 11 occupy aberrant positions. Group 12 occupies an intermediate position. There is a surprisingly smooth progression within Groups 14-22.

The upshot is this: Consider the following sequence of seven phases: {1} {2} {3} {4-5} {6-11} {12-19} {20-22}. Style, as represented by the frequencies of morphemes in List B, varies in a smooth fashion over this sequence of phases. However, it does not vary smoothly within these phases.

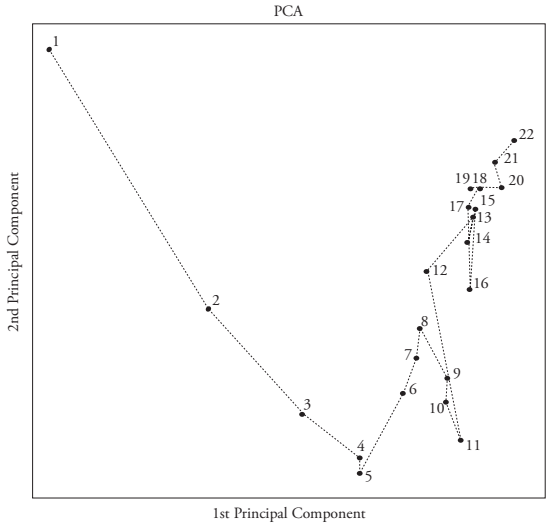

Figure 22. PCA, Groups 1-22, standardized feature vectors, List B. This captures only 36 % of the variance in the data.

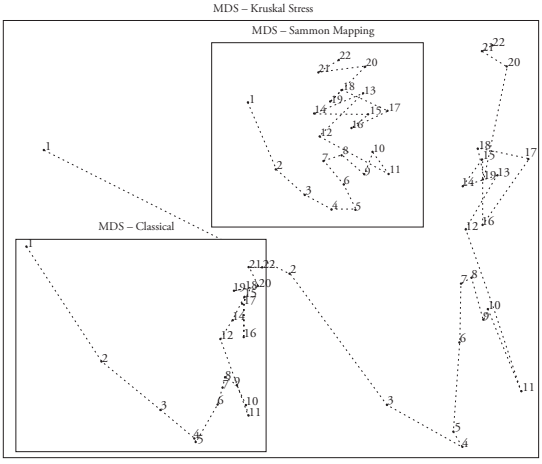

Figure 23. MDS graphs of Groups 1-22, as represented with standardized features from List B. Three methods are used, involving Sammon Mapping (inset, top), Classical MDS (inset, bottom), and Kruskal stress. The stress values are 7.6 % for Kruskal and 8.4 % for Sammon, indicating that the representations are fair. Note how consecutive groups tend to be located near each other. Furthermore, there is an overall progression, with Groups 1 and 22 appearing at the two extremes. Groups 11 and 16 appear at somewhat aberrant locations.

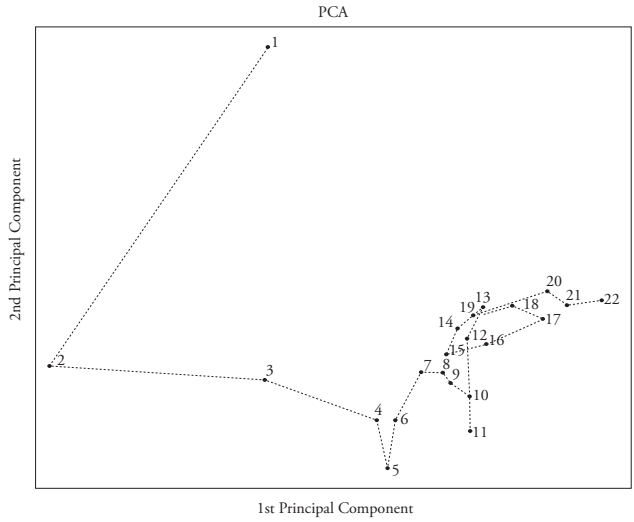

Figure 24. PCA plot of Groups 1-22, as represented with normalized features from List B. The two principal components account for only 36 % of the variance in the data.

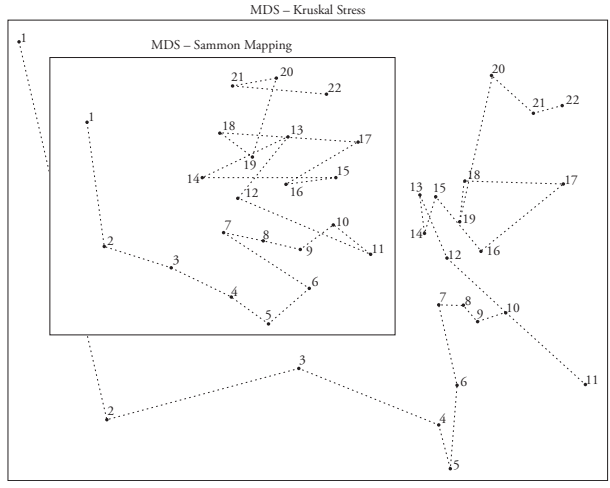

Figure 25. MDS graphs of Groups 1-22, as represented with normalized features from List B. Two methods are used, involving Sammon mapping (inset) and Kruskal stress. The stress values are respectively 8 % and 10 %, indicating that the representations are fair. Other MDS methods lead to similar images.

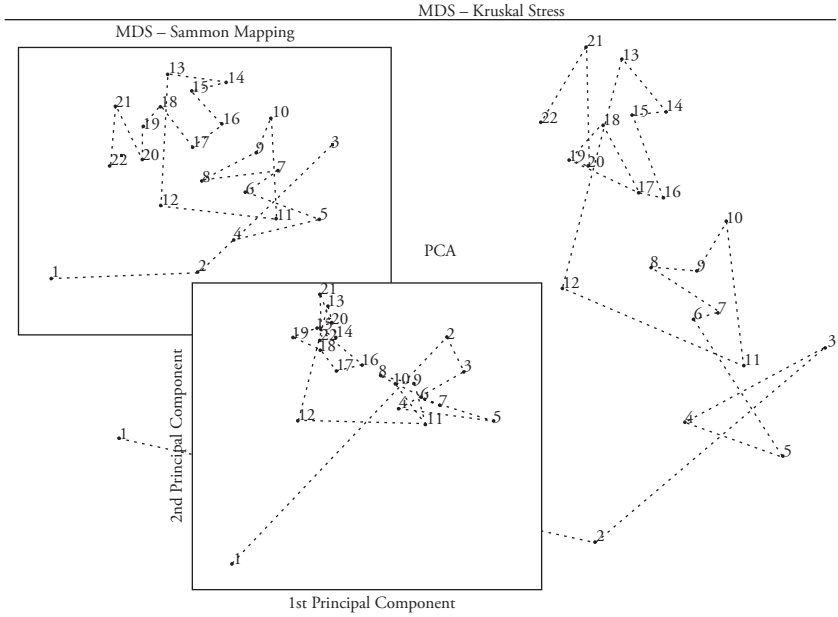

Figure 26. PCA and MDS, Groups 1-22, List B, Optimized weights. The PCA plot (inset, bottom) explains only 56 % of the variance. Sammon mapping (inset top, left) performs fairly well (stress=6.2 %), and MDS-Kruskal performs fairly (stress=12.4 %). Points representing Groups 1-2 must be ignored.

8. Multivariate Analysis (List C): The Generalized Smoking Gun Technique

The multivariate methods employed thus far rely on highly frequent features. Traditional approaches of stylistic analysis, too, may rely on such features, but they do so only in the most extreme cases, *e.g.* where a feature appears in a text with such conspicuous regularity that counting is not necessary. As an example, one may mention the observation of the pre-modern scholars that Meccan verses in the Qurʾān have significantly shorter verse lengths.[58] Stylometry extends this approach, by use of counting, to features whose behavior changes from text to text in a less drastic and noticeable fashion.

[58] For another example, see Behnam Sadeghi, "The Authenticity of Two 2nd/8th-Century Ḥanafī Legal Texts", cited above in footnote 18.

However, more typically, traditional stylistic analysis does not involve the most frequent features. Traditionally, it is more common to rely on the *least* frequent features. That is, if two texts display the same idiosyncrasies, odd spellings, peculiar habits of punctuation, rare words, and unusual phrases, then they are considered to be close in style. This is known as the smoking gun technique. The reason behind its popularity is precisely that it does not involve much computation; a specialist can spot with relative ease, say, any unusual phrases. The main limitation of this method is that many passages may lack a sufficient number of smoking guns.

The smoking gun technique may be transformed using counting and multivariate analysis into a potentially powerful method that one may call *Generalized Smoking Gun*. The generalization involves going beyond rarities, which is the extent of the traditional practice, to include *all* relatively low-frequency words, say, all words that occur fewer than twenty times in the corpus.

This may appear problematic at first. After all, precisely because an infrequent word is infrequent, the variation of its frequency count in two different texts can largely be explained as due to chance. The point, however, is that when one considers several thousand infrequent features simultaneously, the overall patterns may well be highly significant. As the number of features increases, the random oscillations of individual features will tend to cancel each other out, bringing any genuine patterns into bold relief. An advantage of this approach is that it may be applied to passages that are lacking in smoking guns.

Generalized Smoking Gun, however, does have a drawback. It reflects not only style, but also subject matter. In what proportion each factor is represented is unknown to me, requiring tedious analysis to verify, although my sense is that style is no less a factor than subject matter. The same concept may be spoken about using a variety words, and this method captures patterns of word usage by linking passages that tend to use similar vocabulary. Because subject matter also plays a role, one may choose to treat this method as experimental, ignoring it where it contradicts the approaches of the last two sections. But where it agrees with them, there is independent corroboration, as it is highly unlikely for the agreement to be coincidental.

My analysis is based on the list, called List C, of all morphemes that occur in the Qur'ān more than once and fewer than twenty times. There are 3693 such features, and their distribution is shown in Figure 27. The lion's share belongs to morphemes that occur exactly twice in the entire Qur'ān. There is no overlap between this feature list and Lists A and B. So, this feature list offers a genuinely independent marker of style. Each group is represented with a vector of 3693 relative frequency counts—one number for each morpheme. As usual, the task is to investigate the similarities and distances among these vectors.

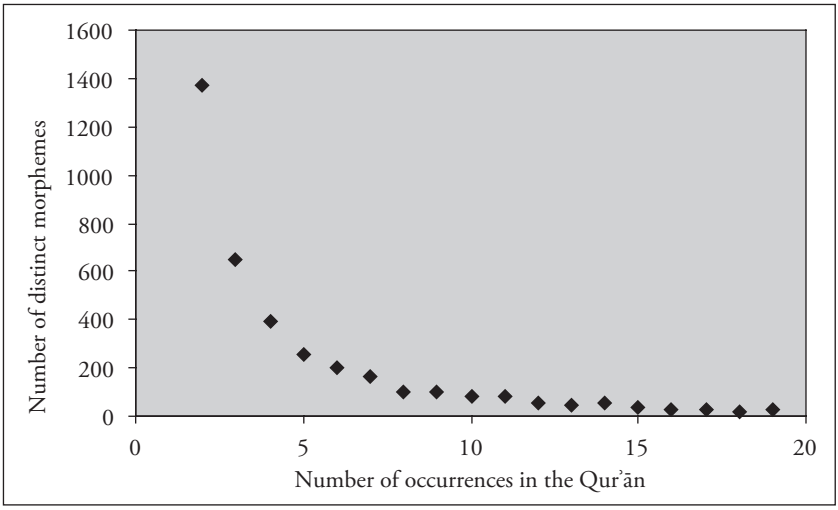

Figure 27. Graph of the number of distinct morphemes that occur a given number of times in the Qurʾān. Thus, 1373 morphemes occur twice, 648 occur three times, and so on. Not included are the morphemes occurring fewer than twice or more than nineteen times.

Figure 28 depicts PCA. The first two principal components explain only fourteen percent of the variance. Normally, this would be considered extremely poor. However, it is different in this case. The dataset is an extremely "noisy" one, being based on thousands of features that, having low frequencies, are highly susceptible to random oscillations in their frequency counts from group to group. In PCA, the noise is accounted for by the principal components that come after the first several ones. Thus much of the variance that the first two components fail to explain is attributable to the large amount of noise that needs to be filtered out anyway. So, unexplained variance is no guide to the quality of the results; but one may fall back on the above-mentioned principle which does provide a measure of objectivity: where the results agree with those from Lists A and B, this has to be significant.

Turning to the PCA graph in Figure 28, one finds impressive agreement with previous results. That is also true of the MDS graphs depicted in Figure 29, which respectively do a "poor" and "fair" job in approximating the distances among the groups. One can see the three regimes familiar from before in both PCA and MDS. Interestingly, the final three groups, 20-22, cluster together at the end of the trajectory as was the case in List B. Group 16 occupies an aberrant position as seen in some previous cases. The main difference with previous cases is that the "leg" corresponding to Regime I (Groups 1-6)

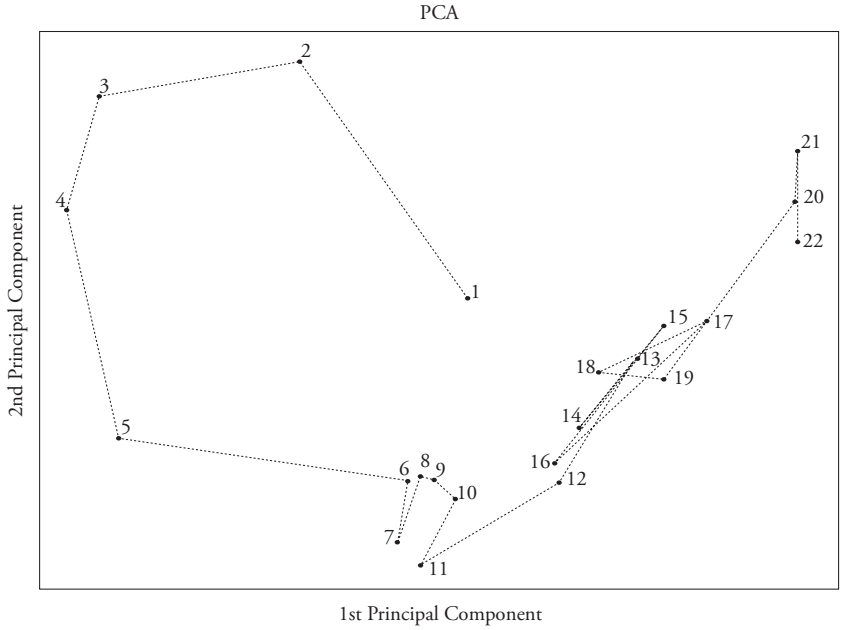

Figure 28. PCA for Groups 1-22, as represented with 3693 infrequent morphemes (those occurring fewer than twenty times in the Qurʾān). The first two principal components explain only 14 % of the variance in the data. The feature counts have been standardized. The normalized case looks more or less the same.

has now folded upon the rest of the trajectory, with Group 5 in the position of the hinge.

Group 1 occupies a central position, and it appears closer to Groups 9 and 14 than Group 2. That means that if Group 1 were joined with these others, we would have a smooth trajectory. But the 1-2-3 sequence, too, looks smooth. The difference is that the latter sequence agrees with the results from Lists A and B and is thus more in keeping with concurrent smoothness.

I have experimented with different ways of weighting the features, for example by giving greater weight to less frequent features, though I have not shown the results here. The results remain about the same, with the main difference lying in the degree to which the "Regime-I leg" folds upon the rest of the trajectory.

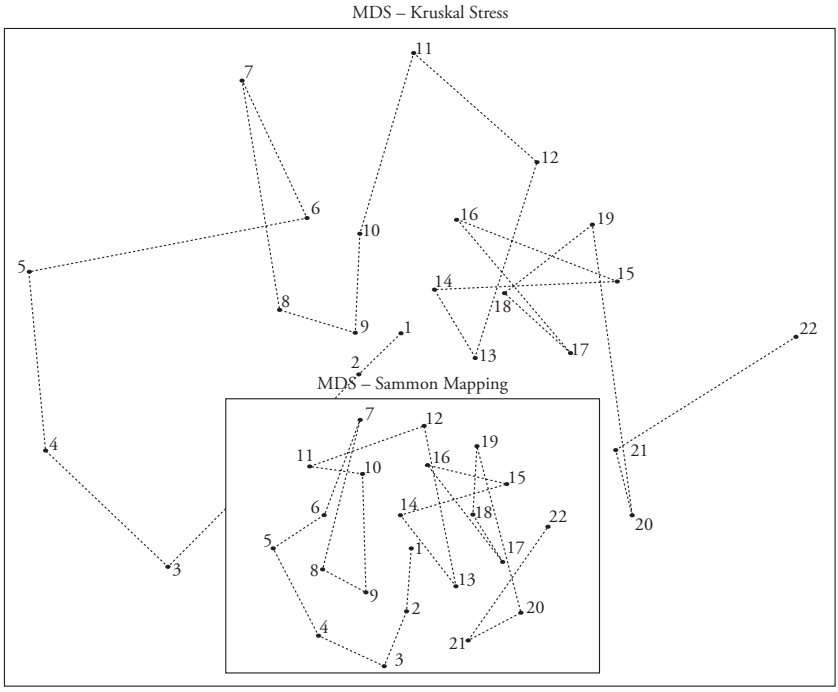

Figure 29. MDS for Groups 1-22, as represented with 3693 infrequent morphemes (those occurring between two to nineteen times in the Qurʾān). The stress values for Sammon (inset) and Kruskal are respectively 12 % (fair to poor) and 20 % (poor). The feature counts have been standardized. The normalized case looks more or less the same.

9. Conclusion

Results

The correlation of the style of the Qurʾān with chronology was recognized in the pre-modern period. Scholars noted that Meccan *sūra*s have shorter verse lengths and employ somewhat different vocabulary and phraseology.[59] Modern scholars have sought to extend this insight beyond the classical binary

[59] See above, footnote 5; cf. Abū l-Faḍl Mīr Muḥammadī, *Buḥūṯ fī tārīḫ al-Qurʾān wa-ʿulūmihi*, Beirut, Dār al-Taʿāruf, 1980, p. 327-36, where he questions the utility of style as evidence for chronology.

Meccan-Medinan division. The Weil-Nöldeke periodization, which was intended as broad and approximate, was based not only on historical and semantic information, but also on style, which was assumed to change monotonically.

Despite its limitations, Bazargan's work remains arguably the most impressive modern attempt at proposing a chronology. Although his criterion of ordering is purely style-based, his corroboration of the resulting sequence came from historical reports and considerations of meaning. My research has provided corroboration for a modified version of his proposed sequence, this time on purely stylistic grounds. The stylometric results of this essay, however, do not merely confirm what we knew already from historical traditions. These traditions, which are not without uncertainties of their own, pin down the dates of a limited number of passages.[60] By contrast, stylometric methods apply to all texts of a certain size, albeit in a statistical sense. They thus provide a broader confirmation than other strands of evidence possibly could.

The following sequence of seven clusters or phases has been corroborated as chronological: {Group 2}, { Group 3}, { Group 4}, { Group 5}, { Groups 6-11}, { Groups 12-19}, { Groups 20-22}. This is the Modified Bazargan chronology (Figure 30).

If one reduces Bazargan's twenty-two groups in this way, then one observes a phenomenon that cannot be due to chance: several different, independent markers of style vary over these phases in a smooth fashion, adjacent clusters having relatively similar stylistic profiles. Such smoothness is observed with *four* markers of style: mean verse length and three powerful, independent

Bazargan	1	2	3	4	5	6-11	12-19	20-22
Modified Bazargan		1	2	3	4	5	6	7

Figure 30. The Modified Bazargan Chronology.

[60] Bāzargān, *Sayr*, vol. I, p. 127-34 / 149-55 / 160-6; Robinson, *Discovering the Qurʾān*, p. 37-44. In addition, the premodern chronological sequences (which can be traced back to a list ascribed to Ibn ʿAbbās) probably constitute early, informed scholarly conclusions and are deserving of serious consideration, but are not necessarily infallible. Moreover, they do not divide *sūra*s into blocks. See *e.g.* Bāzargān, *Sayr*, vol. II, p. 192-203 / 557-69; Robinson, *Discovering the Qurʾān*, p. 284.

multivariate markers. The only discernable explanation for the observed concurrent smoothness is chronological development. One who denies this conclusion has the burden of explaining the pattern in some other way.

There are two exceptions to the consensus of the markers: the distinction between clusters {4} and {5} and that between clusters {12-19} and {20-22} are confirmed by only three of the four markers. The first pair of clusters are blurred together if one considers List B, and the second pair of clusters are lumped together in the case of List A. This makes the distinctions between these pairs of clusters not as well-corroborated as the rest of the sequence.

Bazargan's Group 1 is excluded and no corroboration is claimed for it for three reasons: (1) at 415 words it is not clear whether it is large enough for the chosen markers of style to characterize it meaningfully; (2) its initial position makes it difficult to assess smoothness—and (3) for an important reason to be mentioned later below. Further work is needed on the blocks in Group 1 to determine whether they should be kept together or joined with other blocks/groups.

Univariate markers other than verse length are less powerful, but they nonetheless confirm the order of the initial clusters, as well as the fact that they should appear to the left side of the other clusters.

Granting that the Modified Bazargan sequence reflects chronology, how does one know the direction of the arrow of time? Could it be that the Qur'ān began at the right side of the chain with cluster {20-22} and progressed in reverse of the above sequence, ending at cluster {2}? Actually, stylistic evidence does not settle this issue at all. Instead, the question is answered by considerations of meaning. For example, Block 110 is an *exegetical* insertion in *sūra* 74, explaining a point in Block 27. As such, Block 110 (in Group 7) presupposes the existence of Block 27 (Group 2) and is therefore later.[61] Such an argument does not take meaning into consideration except in the minimal manner of noting that a sentence obviously clarifies and presupposes another one. But if one were willing to go deeper into meaning, a variety of other indications could be marshalled to support the same direction. For example, passages referring to oppression as something of the past must be placed after passages responding to ongoing oppression.

Now that we know head from tail, it is of interest to comment on the rate of stylistic change. If one accepts the broad outlines of the traditional reckoning of chronology and the division into Meccan and Medinan periods, and if

[61] On this insertion, see Bāzargān, *Sayr*, vol. II, p. 150-2 / 543; Angelika Neuwirth, *Studien zur Komposition der mekkanischen Suren*, Berlin, Walter de Gruyter, 1981, p. 215; Tilman Nagel, *Medinensische Einschübe in mekkanischen Suren*, Göttingen,Vandenhoeck & Ruprecht, 1995, p. 89. For the definitions of the blocks, see above, Table 1 and Table 2.

one makes the heuristic assumption that the text was disseminated at a roughly even rate, then from both univariate and multivariate markers, one discerns that style changed rapidly at the beginning. The pace of change slowed gradually. The initially more rapid pace grants style greater discriminatory power in the earlier phases. It is for this reason that even the relatively weaker univariate methods have no problem detecting the initial eruption. It is also for this reason that the Meccan period claimed three of the four phases of Weil's chronology. After {Group 5}, however, univariate markers at best give a vague sense of continuity, failing to identify the patterns detected by multivariate methods. Incidentally, the greater discriminatory power of style for the Meccan period is fortunate, as in this period historical indications of chronology are sparse by comparison. Meccan texts contain fewer explicit references to "current events" than Medinan ones.

It should be stressed that this chronology should not be treated in a rigid way. The fact that cluster {6-11} comes after {4} does not mean that every passage in {4} is in fact older than everything in {6-11}. The sequence is valid in an *average* sense only. Deviations from averages, as well as outlier behavior, are typical for phenomena complex enough to merit statistical analysis. One can expect that to be the case here as well. Of course, a common goal in statistics is estimating the likelihood of deviations of a certain size from the average. Establishing such confidence intervals and increasing precision are long-term goals of Qurʾānic stylometry.

What about the *internal* chronologies of the three clusters with more than one group: {6-11}, {12-19}, and {20-22}? Because I have failed to establish concurrent smoothness internally within these clusters, their internal chronologies are indeterminate. Is Bazargan right that Group 19 came after 18, which came after 17, and so on through Group 12? The possibilities are threefold: (*i*) He is right, but in this period style did not in reality change over time concurrently smoothly, or (*ii*) it did, and he is right, but the methods employed in this essay are not sensitive enough to reveal it consistently, or *(iii)* Bazargan's internal chronology of these clusters is not correct. There are some indications for *iii*, such as the fact that independent multivariate markers of style consistently assign Group 16 to an earlier time than Bazargan does, thus indicating that here a sequence different from that of Bazargan is likely to satisfy concurrent smoothness. (Perhaps not coincidentally, the largest texts in Group 16 are Blocks 155 and 157, both of which are traditionally considered as Meccan.) A combination of these three possibilities could hold as well. More research is needed for better answers to these questions.

If the reader is disappointed with the limitations thus far, it might cheer him or her to know that, as stylistic studies of chronology go, the Qurʾān

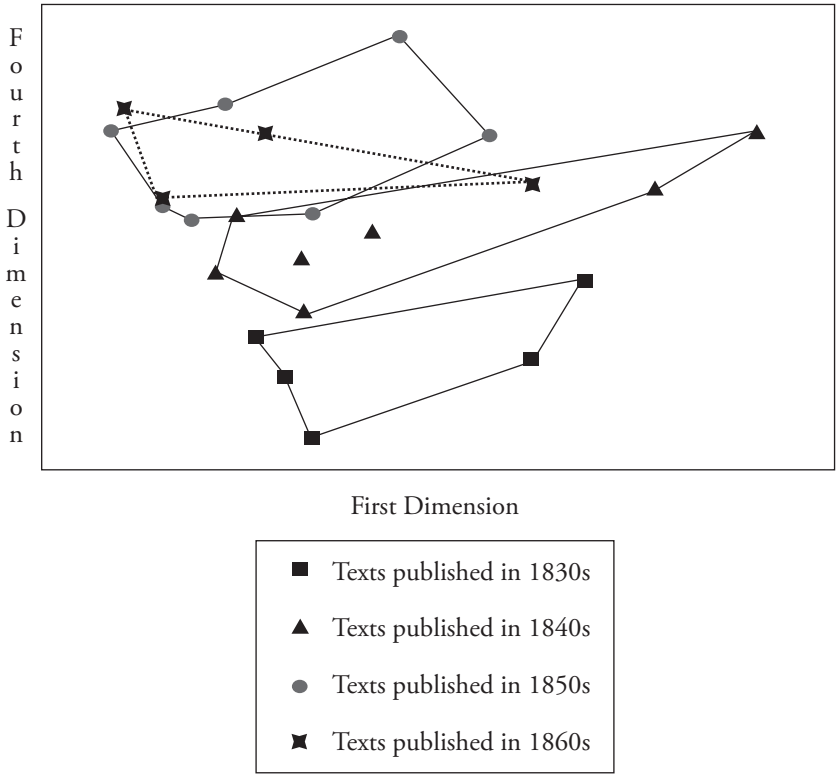

Figure 31. Tabata's application of a method called ANACOR, which is analogous to PCA, to twenty-three works of Charles Dickens. Three consecutive chronological clusters emerge along the first and the fourth dimensions (analogous to the first and fourth principal components in PCA), arranged vertically in the diagram. Tabata concludes, "with a sustained change in style from the 1830s through the 1840s, Dickens seems to have established his style by about 1850". (Tabata, "Investigating stylistic variation in Dickens", p. 178.)

yields impressive results. For example, after over a century of stylometric studies, Plato's dialogues have not offered conclusions of similar force, clarity, and precision.

As another example, it is worthwhile to reflect on Tomoji Tabata's elegant study of Charles Dickens.[62] Figure 31 depicts twenty-three works of Dickens spanning four decades using multivariate analysis of thirty-four word-class

[62] Tabata, "Investigating Stylistic Variation in Dickens", cited in footnote 18.

variables (*e.g.* singular nouns, possessive nouns, etc.). One observes no more than three consecutive clusters. The first two clusters correspond respectively to the first and second decades of Dickens' carrier while the third cluster covers the remaining two decades.

The level of discrimination achieved in the study of the Qurʾān is all the more remarkable considering several serious actual and potential sources of error:

First, the division of *sūra*s into blocks will always include an element of conjecture, making it likely that there are errors. Such errors cause the attrition of smoothness, thus reducing the number of phases that can be obtained that display concurrent smoothness.

Second, the blocks are very small. Ninety blocks contain fewer than 300 words apiece! To obtain larger samples, one must combine many blocks into groups. This may be done according to different criteria, such as verse length profiles as in Bazargan's work. Regardless of the method used to combine blocks, however, one can expect that some blocks will be assigned together that are not in reality from the same period. This again will diminish smoothness and reduce the number of phases that can exhibit concurrent smoothness.

Third, the Qurʾān as a whole constitutes a relatively small corpus. The twenty-two groups add up to approximately 78,000 word-tokens. By comparison, the above-mentioned Dickensian texts contain about 4.7 million word-tokens. All but three of the individual works in that corpus are larger than the entire Qurʾān. A smaller text typically means lesser statistical significance and smaller chances for devising experimental controls.

Fourth, while the Qurʾān displays continuous generic evolution, there may also be room for speaking of discrete genres in the Qurʾān, specifically with regard to legal vs. non-legal material. Studies have shown that discrete genres can affect style heavily. For better results, the variables of time and discrete genre must be studied together.

This omission, by the way, does not invalidate the degree of discrimination that has been achieved. The observed patterns cannot be explained in terms of differences in the proportions of *discrete* styles (genres) used.[63] A plausible

[63] Against my conclusions, one might argue that the stylistic variation over the seven clusters merely reflects that each cluster has material of different *discrete* stylistic profiles (*e.g.* genres) in differing proportions. For example, {2} would have no legal content, {3} would have a little, {4} would have a little more, and so on. Although the gradual decline in the amount of eschatological material in the Meccan phase (shown in the graphs in Bāzargān, *Sayr*, vol. I, p. 177-8 / 201-2 / missing) may have contributed somewhat to such an effect, on the whole what one observes is genuinely *gradual* change in style rather than the mixing of *discrete* styles in gradually changing proportions; thus eschatological passages experience stylistic evolution as well. Gener-

exception to this would be in the case of Groups 21 and 22, which have a much higher proportion of legal material than other groups (up to a third). One must investigate (*i*) whether legal material indeed comprises a very distinct stylistic profile, and (*ii*) if yes, how much this has contributed to the clustering of Groups 20-22. A similar case might be made for Group 1, which largely consists of what may be called "introductory sections", i.e. the beginning portions of *sūra*s. It is possible that Meccan introductory sections are characterized by a distinctive stylistic register. This is the third reason for my having excluded Group 1.

Finally, I need to address a potential problem regarding the possibility of discontinuity in style. Could it be that at some point in time Qurʾānic revelation breaks with its current style, reverts to a much earlier style for a while, and then leaps forward to resume where it left off? If so, what would that imply for stylistic analysis? Note that this scenario does not *merely* envisage stylistic discontinuity. In general, discontinuity *per se* would not undermine my approach: it would simply mean that one cannot observe concurrent smoothness, leading one to conclude nothing (as opposed to concluding something that is wrong). Rather, this scenario poses a particular form of discontinuity, one involving a complete reversion to a previous style, comprising all markers (vocabulary, grammatical structures, verse length, word length, etc.). If this sort of thing did happen, then the outcome would be two passages from different periods with *exactly* the same style. Not able to differentiate between them, stylistic analysis could incorrectly assign them to the same time. Crucially, this would not entail a perturbation of concurrent smoothness.

Research into whether this happened is needed. Even after such research, the scenario probably cannot be eliminated completely and could still apply to some passages. To the extent that this possibility remains open, I accept it as a potential source of error in my results. Nevertheless, I do not think that this scenario is probable, or that it could have happened on a large scale. It is odd to think that the style could leap back to that of a specific bygone year in all its particulars rather than just in some features. Furthermore, the scenario would have been more plausible if stylistic evolution in the Qurʾān were characterized by the employment of discrete styles (*e.g.* corresponding to discrete genres). Then it would not be odd if the Qurʾān switched to a formerly more common discrete style (corresponding to a specific genre) and then back. In reality, however, style tends to form a continuum.[64] Moreover, it does so even within a genre. Specific genres (such as eschatological and legal) evince evolution in style as well, although more research should be done on this point.

---

ally, the Qurʾān's style comprises a continuum rather than sharp, discrete categories.

[64] See the last footnote.

Additionally, tests performed in Section 5 support the centrality of time as a variable.

Implications

Literary sources and manuscript evidence indicate that the Prophet Muḥammad disseminated the contents of the Qurʾān, and that the Caliph ʿUṯmān dispatched master copies of the scripture to several cities.[65] As a *thought experiment*, however, let us unlearn what we know, imagining that we had come across the Qurʾān not knowing where it came from or who disseminated it. In fact, let us even overlook the semantic contents of the text. What could one conclude about the Qurʾān's composition just from the formal-stylistic patterns observed?

One would conclude that *style backs the hypothesis of one author*. For the sake of argument, suppose there were two authors: let's say *A* wrote Groups 1-11, and *B* wrote 12-22. Then one would have to say that the style of *A* moved along a trajectory towards that of *B*, or that *B* picked up where *A* stopped, with respect to not only verse length, but also frequencies of the most common morphemes and frequencies of uncommon words. It is much easier to imagine a single author. Furthermore, if one assumed three, four, or more authors instead of two, the improbability would increase exponentially. Imagine three authors: author *A* being responsible for Groups {1-5}, *B* for {6-11}, and *C* for {12-22}. Again, one would have to explain why *A*'s style gravitated over time towards that of *B*, not only in terms of verse length, but also in terms of vocabulary usage for high-, medium-, and low-frequency morphemes. Moreover, one would need to explain why the style of *B* forms an intermediate stage between those of *A* and *C* in terms of the various markers of style that have been considered. If one imagined seven authors, one would have to explain why each person's style is between two others not only with respect to verse length, but also with respect to frequent morphemes, medium-frequency morphemes, and unusual words.

The study reveals the stylistic continuity and distinctiveness of the text as a whole. As far as this point is concerned, the present study makes palpable what we knew already: no competent and seasoned scholar of the Qurʾān, while aware of the stylistic variation in the text, could lose sight of its underlying unity.

Also established now is the general integrity of many passages of long and medium size. That goes against Richard Bell's instincts. His thought-provoking vision, while not implausible historically, now appears to have been

[65] For a discussion of some of the evidence, see Behnam Sadeghi, "The Codex of a Companion of the Prophet and the Qurʾān of the Prophet", cited above in footnote 4.

misguided. Not only do eighteen test cases examined display a surprising degree of stylistic unity (Section 5), but also on his viewpoint one would be hard-pressed to explain the degree of concurrent smoothness observed.[66] Moreover, one would expect revisions of earlier passages to dilute their stylistic distinctiveness. In fact, while it cannot be denied that the Prophet revised some passages over time, the present study shows that such revisions could not have been extensive.

There are also implications for the *Sīra* literature. It has always been evident that there is a fit between the Prophet's biography in the *Sīra* and the Qurʾān's style and contents. At the broadest level, the *hiǧra* divides the Prophet's career into two different periods, a nice fit with the fact that stylistically these periods are distinct. More particularly, there are apparent links between the major events of the Prophet's career and specific passages. The connections are noteworthy enough for the *Sīra* to offer a "plausible chronological framework for the revelations".[67] These connections have normally been exploited to shed light on the chronology of the Qurʾān. For example, Bazargan uses them to test and calibrate his chronology. Thus, the flow of information has been more often from the *Sīra* to the Qurʾān, although one could also have argued for mutual corroboration between the *Sīra* and the Qurʾān. Now, however, that one is able in some measure to evaluate chronologies of the Qurʾān without resort to historical reports, one can reverse the direction of information flow: to the extent that a style-supported chronology fits the *Sīra*'s outline, there is independent substantiation of the *Sīra*.

Directions for Future Research

The main way in which Bazargan's religious faith affected his scholarship was to make it more critical and rigorous. Conscious of the gravity of mischaracterizing the "Words of God", he reasoned cautiously, emphasized the fallible nature of his work, describing it as a research program "in its infancy", and invited other scholars to critique, correct, and refine his findings.[68] In a similar spirit, I have done my best to highlight the known limitations and potential shortcomings of my approach. However, at least one thing is amply clear by now: the utility of stylistic analysis as an effective means of determining the

[66] Using methods different from mine, Angelika Neuwirth and Neal Robinson, too, reach conclusions that are incompatible with Bell's approach (Neuwirth, *Studien*; Robinson, *Discovering the Qurʾān*, p. 94-5, 162, 177-84, 187, 191). Although Bell is almost always wrong, his observations on particular points are sometimes worth thinking about.

[67] Robinson, *Discovering the Qurʾān*, p. 37.

[68] Bāzargān, *Sayr Mutammim*, p. 409; *Sayr*, vol. II, p. v / 195-7.

chronology of the Qurʾān cannot be denied. Scholars would ignore style at their peril.

Much more remains to be done on chronology in several different areas of research: the information in *ḥadīṯ*s and the *Sīra*, traditional stylistic analysis, and stylometry. As far as stylometry is concerned, what has been done here scratches the surface of what can be attempted. It may be possible to develop yet more effective markers of style, particularly by combining the frequency of syntactical and morphological features with the different types of features used so far. Furthermore, one may refine the method of weight optimization by using more training texts and performing separate optimizations for different phases (*e.g.* Meccan and Medinan). Studying the variable of genre and its interaction with the variable of time remains a desideratum. Most significantly, constructing a better chronology is within reach. As a discriminator of chronology, verse length loses much of its efficacy in the Medinan period.[69] One must therefore also use other markers of style to divide and reorder the text. The new sequences obtained should be examined for concurrent smoothness. The ultimate goal is to maximize the number of phases that display concurrent smoothness. Research in that area is in progress.

[69] See the analysis in "Univariate Marker of Style: Mean Verse Length", above in Section 3, especially Table 4, Figure 5, and Figure 6.

11. Appendix

List A

Table 10. ***Distance matrix for Groups 1-22, Feature List A, normalized: table of pairwise distances among the groups. These are city-block distances, rescaled so that the maximum distance is 1,000.***

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22
1	0	649	759	780	865	966	927	831	853	956	897	872	912	916	933	1000	943	976	954	982	877	916
2	649	0	488	477	529	477	433	436	494	596	489	573	616	548	663	686	677	574	602	559	618	599
3	759	488	0	301	370	484	498	469	471	456	453	596	662	585	584	560	600	626	639	666	631	647
4	780	477	301	0	231	360	430	369	365	381	386	548	564	516	552	475	579	534	536	560	592	629
5	865	529	370	231	0	324	443	442	369	346	412	501	591	500	531	452	523	556	529	545	600	609
6	966	477	484	360	324	0	282	330	369	309	287	456	528	458	504	389	473	369	429	417	554	514
7	927	433	498	430	443	282	0	265	301	331	332	364	455	366	481	376	440	355	395	363	445	387
8	831	436	469	369	442	330	265	0	298	273	375	360	391	289	370	307	337	315	280	360	409	415
9	853	494	471	365	369	369	301	298	0	255	286	252	381	294	363	323	335	365	392	391	375	326
10	956	596	456	381	346	309	331	273	255	0	325	286	367	318	298	223	326	287	314	407	406	394
11	897	489	453	386	412	287	332	375	286	325	0	388	526	473	470	423	494	447	480	497	510	418
12	872	573	596	548	501	456	364	360	252	286	388	0	348	285	282	292	328	345	337	406	370	286
13	912	616	662	564	591	528	455	391	381	367	526	348	0	253	243	341	354	344	247	301	223	304
14	916	548	585	516	500	458	366	289	294	318	473	285	253	0	239	327	281	266	217	322	297	314
15	933	663	584	552	531	504	481	370	363	298	470	282	243	239	0	292	297	259	242	351	278	299
16	1000	686	560	475	452	389	376	307	323	223	423	292	341	327	292	0	303	311	265	401	356	378
17	943	677	600	579	523	473	440	337	335	326	494	328	354	281	297	303	0	285	290	332	259	278
18	976	574	626	534	556	369	355	315	365	287	447	345	344	266	259	311	285	0	231	264	298	317
19	954	602	639	536	529	429	395	280	392	314	480	337	247	217	242	265	290	231	0	286	262	290
20	982	559	666	560	545	417	363	360	391	407	497	406	301	322	351	401	332	264	286	0	240	292
21	877	618	631	592	600	554	445	409	375	406	510	370	223	297	278	356	259	298	262	240	0	196
22	916	599	647	629	609	514	387	415	326	394	418	286	304	314	299	378	278	317	290	292	196	0

Table 11. ***Distance matrix for Groups 3-22, Feature List A, optimized weights. These are city-block distances, rescaled so that the maximum distance is 1,000.***

	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22
3	0	406	540	784	712	642	580	549	522	665	680	637	629	641	707	1000	823	963	826	820
4	406	0	409	547	621	440	394	601	381	615	611	463	585	530	709	758	625	718	716	718
5	540	409	0	534	547	497	408	384	478	526	655	495	596	535	597	788	618	730	742	709
6	784	547	534	0	312	444	461	541	436	524	631	535	600	479	607	440	495	427	744	607
7	712	621	547	312	0	419	427	607	493	419	593	480	753	524	613	629	577	417	678	554
8	642	440	497	444	419	0	249	413	452	273	451	275	419	332	409	497	312	448	533	471
9	580	394	408	461	427	249	0	388	340	221	369	245	411	288	361	624	347	555	450	383
10	549	601	384	541	607	413	388	0	401	335	563	550	390	359	385	602	484	712	641	543
11	522	381	478	436	493	452	340	401	0	424	499	457	461	284	581	623	515	646	587	494
12	665	615	526	524	419	273	221	335	424	0	346	299	362	252	286	570	310	532	403	283
13	680	611	655	631	593	451	369	563	499	346	0	270	306	407	423	605	323	419	214	246
14	637	463	495	535	480	275	245	550	457	299	270	0	356	357	416	533	234	419	388	334
15	629	585	596	600	753	419	411	390	461	362	306	356	0	296	368	453	249	502	382	319
16	641	530	535	479	524	332	288	359	284	252	407	357	296	0	347	482	312	495	414	349
17	707	709	597	607	613	409	361	385	581	286	423	416	368	347	0	498	320	581	425	422
18	1000	758	788	440	629	497	624	602	623	570	605	533	453	482	498	0	345	356	509	454
19	823	625	618	495	577	312	347	484	515	310	323	234	249	312	320	345	0	349	344	262
20	963	718	730	427	417	448	555	712	646	532	419	419	502	495	581	356	349	0	370	339
21	826	716	742	744	678	533	450	641	587	403	214	388	382	414	425	509	344	370	0	191
22	820	718	709	607	554	471	383	543	494	283	246	334	319	349	422	454	262	339	191	0

Table 12. ***Nearest Neighbors, List A. These tables were devised using the distance matrices and offer a convenient way of interpreting them. The groups are listed consecutively in the leftmost columns. The second column gives the nearest groups to the leftmost ones, the third column gives the second nearest groups to the leftmost ones, and so on. To gain an idea of the neighborhood of a group, two steps must be taken. First, find the group in the leftmost column and look up its nearest neighbors. Second, find the group in the other columns and look up the two leftmost groups.***

	Normalized features				
1	2	3	4	8	9
2	7	8	4	6	3
3	4	5	11	10	8
4	5	3	6	9	8
5	4	6	10	9	3
6	7	11	10	5	8
7	8	6	9	10	11
8	7	10	19	14	9
9	12	10	11	14	8
10	16	9	8	12	18
11	9	6	10	7	8
12	9	15	14	10	22
13	21	15	19	14	20
14	19	15	13	18	17
15	14	19	13	18	21
16	10	19	12	15	17
17	21	22	14	18	19
18	19	15	20	14	17
19	14	18	15	13	21
20	21	18	19	22	13
21	22	13	20	17	19
22	21	17	12	19	20

	Standardized features				
1	2	3	4	12	21
2	8	7	3	11	6
3	4	5	11	10	9
4	5	3	9	8	6
5	4	10	9	3	6
6	10	7	11	8	5
7	8	9	11	6	10
8	10	19	7	14	16
9	10	12	11	14	7
10	8	16	9	18	6
11	9	7	6	10	8
12	15	9	16	14	10
13	15	21	19	14	22
14	15	19	8	13	9
15	19	13	14	12	18
16	19	10	15	8	12
17	18	19	21	22	15
18	19	15	17	20	10
19	18	15	16	14	21
20	21	18	19	13	22
21	22	13	20	19	15
22	21	19	13	20	17

	Weight-optimized features				
3	4	11	5	10	9
4	11	9	3	5	8
5	10	9	4	11	14
6	7	20	11	18	8
7	6	20	8	12	9
8	9	12	14	19	16
9	12	14	8	16	11
10	12	16	5	17	9
11	16	9	4	10	12
12	9	16	8	22	17
13	21	22	14	15	19
14	19	9	13	8	12
15	19	16	13	22	14
16	12	11	9	15	19
17	12	19	16	9	15
18	19	20	6	15	22
19	14	15	22	12	8
20	22	19	18	21	7
21	22	13	19	20	15
22	21	13	19	12	15

List B

Table 13. ***Distance Matrix, Feature List B, standardized: table of pairwise distances among the groups. These are city-block distances, rescaled so that the maximum distance is 1,000.***

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22
1	0	689	807	873	841	859	845	856	971	937	1000	804	858	846	880	842	885	820	780	922	918	933
2	689	0	562	690	741	667	651	703	750	749	838	646	776	696	732	708	741	779	724	831	761	820
3	807	562	0	580	520	597	520	567	598	568	646	560	607	579	614	582	622	631	640	681	616	716
4	873	690	580	0	473	519	519	485	511	494	533	493	613	598	588	557	640	644	581	643	708	674
5	841	741	520	473	0	450	474	474	517	506	581	516	603	623	633	529	616	621	614	689	686	669
6	859	667	597	519	450	0	461	457	492	504	533	476	534	537	533	483	587	515	533	621	630	647
7	845	651	520	519	474	461	0	424	474	465	528	435	487	493	539	423	564	507	468	573	598	573
8	856	703	567	485	474	457	424	0	457	441	523	442	530	475	497	456	555	497	479	606	595	568
9	971	750	598	511	517	492	474	457	0	474	576	461	573	523	529	501	611	522	542	638	637	613
10	937	749	568	494	506	504	465	441	474	0	535	460	542	526	540	433	583	531	509	613	664	623
11	1000	838	646	533	581	533	528	523	576	535	0	579	599	577	639	550	615	616	592	681	609	666
12	804	646	560	493	516	476	435	442	461	460	579	0	498	452	439	459	576	459	475	548	531	540
13	858	776	607	613	603	534	487	530	573	542	599	498	0	502	455	408	533	446	434	547	529	605
14	846	696	579	598	623	537	493	475	523	526	577	452	502	0	435	408	522	438	491	601	552	510
15	880	732	614	588	633	533	539	497	529	540	639	439	455	435	0	492	490	468	451	537	529	529
16	842	708	582	557	529	483	423	456	501	433	550	459	408	408	492	0	498	446	433	519	534	531
17	885	741	622	640	616	587	564	555	611	583	615	576	533	522	490	498	0	562	527	584	522	560
18	820	779	631	644	621	515	507	497	522	531	616	459	446	438	468	446	562	0	408	486	544	466
19	780	724	640	581	614	533	468	479	542	509	592	475	434	491	451	433	527	408	0	513	529	525
20	922	831	681	643	689	621	573	606	638	613	681	548	547	601	537	519	584	486	513	0	493	507
21	918	761	616	708	686	630	598	595	637	664	609	531	529	552	529	534	522	544	529	493	0	467
22	933	820	716	674	669	647	573	568	613	623	666	540	605	510	529	531	560	466	525	507	467	0

Table 14. ***Distance matrix for Groups 3-22, Feature List B, optimized weights. These are city-block distances, rescaled so that the maximum distance is 1,000.***

	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22
3	0	791	617	766	754	719	900	800	689	759	804	768	836	812	794	952	885	909	765	797
4	791	0	773	675	621	647	756	513	522	645	723	593	687	699	733	857	668	790	856	694
5	617	773	0	745	873	824	940	861	726	838	923	846	892	829	815	1000	893	958	950	839
6	766	675	745	0	620	676	534	779	599	531	530	694	511	687	572	567	576	616	737	650
7	754	621	873	620	0	477	651	638	594	482	533	532	694	583	807	618	482	619	767	541
8	719	647	824	676	477	0	706	617	529	576	605	549	774	607	857	760	665	761	706	652
9	900	756	940	534	651	706	0	744	751	586	693	675	735	608	801	750	694	797	824	685
10	800	513	861	779	638	617	744	0	640	714	784	624	779	575	857	837	629	819	973	749
11	689	522	726	599	594	529	751	640	0	629	645	554	697	670	671	762	631	780	753	605
12	759	645	838	531	482	576	586	714	629	0	406	513	520	505	651	596	459	563	575	574
13	804	723	923	530	533	605	693	784	645	406	0	546	513	570	666	597	533	644	652	702
14	768	593	846	694	532	549	675	624	554	513	546	0	770	459	731	622	515	678	675	572
15	836	687	892	511	694	774	735	779	697	520	513	770	0	813	559	781	620	718	781	707
16	812	699	829	687	583	607	608	575	670	505	570	459	813	0	785	624	495	649	706	613
17	794	733	815	572	807	857	801	857	671	651	666	731	559	785	0	731	681	737	746	632
18	952	857	1000	567	618	760	750	837	762	596	597	622	781	624	731	0	474	494	764	565
19	885	668	893	576	482	665	694	629	631	459	533	515	620	495	681	474	0	565	805	563
20	909	790	958	616	619	761	797	819	780	563	644	678	718	649	737	494	565	0	828	622
21	765	856	950	737	767	706	824	973	753	575	652	675	781	706	746	764	805	828	0	539
22	797	694	839	650	541	652	685	749	605	574	702	572	707	613	632	565	563	622	539	0

Table 15. ***Nearest Neighbors, List B. These tables were drawn up using the distance matrices and offer a convenient way of interpreting them. Just follow the instructions in the caption for Table 12.***

	Normalized features					Standardized features						Weight-optimized features					
1	19	14	12	18	2	1	2	19	12	3	18						
2	3	7	14	12	4	2	3	12	7	6	1						
3	7	8	5	12	14	3	5	7	12	2	8	3	5	11	8	7	12
4	8	12	5	7	9	4	5	8	12	10	9	4	10	11	14	7	12
5	4	7	6	8	12	5	6	4	7	8	10	5	3	11	6	4	17
6	8	7	5	12	9	6	5	8	7	12	16	6	15	13	12	9	18
7	8	16	12	19	9	7	16	8	12	6	10	7	8	12	19	14	13
8	7	12	9	10	16	8	7	10	12	16	6	8	7	11	14	12	13
9	12	8	7	10	16	9	8	12	7	10	6	9	6	12	16	7	14
10	8	16	12	7	9	10	16	8	12	7	9	10	4	16	8	14	19
11	8	7	10	4	6	11	8	7	4	6	10	11	4	8	14	7	6
12	16	7	19	14	9	12	7	15	8	14	16	12	13	19	7	16	14
13	16	19	18	7	12	13	16	19	18	15	7	13	12	15	6	7	19
14	16	12	15	8	18	14	16	15	18	12	8	14	16	12	19	7	13
15	12	14	19	13	8	15	14	12	19	13	18	15	6	13	12	17	19
16	14	13	19	7	12	16	13	14	7	10	19	16	14	19	12	13	10
17	16	21	19	14	22	17	15	16	14	21	19	17	15	6	22	12	13
18	19	13	12	16	14	18	19	14	13	16	12	18	19	20	22	6	12
19	18	16	13	12	7	19	18	16	13	15	7	19	12	18	7	16	14
20	21	18	19	16	22	20	18	21	22	19	16	20	18	12	19	6	7
21	22	20	17	19	18	21	22	20	17	13	15	21	22	12	13	14	8
22	21	18	16	19	17	22	18	21	20	14	19	22	21	7	19	18	14

Table 16. ***Distance Matrix, Feature List C, standardized: table of pairwise distances among the groups. These are city-block distances, rescaled so that the maximum distance is 1,000.***

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22
1	0	465	660	660	681	596	681	629	585	608	648	620	574	511	611	605	586	594	609	665	636	667
2	465	0	789	802	828	774	835	795	770	796	828	803	761	694	800	784	775	774	790	857	815	851
3	660	789	0	866	865	880	953	869	870	877	920	921	878	854	923	912	908	886	922	973	958	975
4	660	802	866	0	868	869	923	887	837	873	918	903	879	854	922	884	925	901	934	983	951	1000
5	681	828	865	868	0	828	884	864	848	856	910	895	890	843	919	874	919	880	918	988	962	991
6	596	774	880	869	828	0	835	813	804	785	806	811	821	744	819	795	816	810	824	865	864	892
7	681	835	953	923	884	835	0	849	846	810	875	881	845	805	885	832	886	869	865	944	931	919
8	629	795	869	887	864	813	849	0	777	784	853	829	792	761	831	805	817	795	830	893	875	911
9	585	770	870	837	848	804	846	777	0	735	831	783	771	707	835	762	793	782	820	870	846	875
10	608	796	877	873	856	785	810	784	735	0	800	819	785	730	801	768	799	776	814	863	864	888
11	648	828	920	918	910	806	875	853	831	800	0	842	838	792	869	833	860	833	862	927	910	921
12	620	803	921	903	895	811	881	829	783	819	842	0	763	753	814	797	818	805	797	878	867	875
13	574	761	878	879	890	821	845	792	771	785	838	763	0	701	742	752	765	757	758	830	795	831
14	511	694	854	854	843	744	805	761	707	730	792	753	701	0	752	742	736	731	730	816	776	800
15	611	800	923	922	919	819	885	831	835	801	869	814	742	752	0	793	743	784	789	849	802	848
16	605	784	912	884	874	795	832	805	762	768	833	797	752	742	793	0	772	747	786	851	831	868
17	586	775	908	925	919	816	886	817	793	799	860	818	765	736	743	772	0	763	793	839	790	828
18	594	774	886	901	880	810	869	795	782	776	833	805	757	731	784	747	763	0	793	827	822	815
19	609	790	922	934	918	824	865	830	820	814	862	797	758	730	789	786	793	793	0	852	817	840
20	665	857	973	983	988	865	944	893	870	863	927	878	830	816	849	851	839	827	852	0	810	863
21	636	815	958	951	962	864	931	875	846	864	910	867	795	776	802	831	790	822	817	810	0	845
22	667	851	975	1000	991	892	919	911	875	888	921	875	831	800	848	868	828	815	840	863	845	0

Table 17. ***Nearest Neighbors, List C. These tables were drawn up using the distance matrices. See the caption for Table 12.***

	Normalized features					Standardized features					
1	14	2	13	9	16	1	2	14	13	9	17
2	1	14	13	9	16	2	1	14	13	9	6
3	1	2	14	9	8	3	1	2	14	5	4
4	1	14	9	2	13	4	1	2	9	14	3
5	1	14	9	6	10	5	1	2	6	14	9
6	1	14	10	16	9	6	1	14	2	10	16
7	1	14	10	16	9	7	1	14	10	16	2
8	1	14	9	10	13	8	1	14	9	10	13
9	1	14	10	16	13	9	1	14	10	16	2
10	1	14	9	16	13	10	1	14	9	16	18
11	1	14	10	6	16	11	1	14	10	6	2
12	1	14	13	9	16	12	1	14	13	9	16
13	1	14	16	15	9	13	1	14	15	16	18
14	1	13	9	2	10	14	1	2	13	9	10
15	1	14	13	17	16	15	1	13	17	14	18
16	1	14	13	18	9	16	1	14	18	13	9
17	1	14	13	15	16	17	1	14	15	18	13
18	1	14	16	13	17	18	1	14	16	13	17
19	1	14	13	16	15	19	1	14	13	16	15
20	1	14	13	18	21	20	1	21	14	18	13
21	1	14	13	17	15	21	1	14	17	13	15
22	1	14	18	13	17	22	1	14	18	17	13

Precise Definition for Smoothness

Given an ordered sequence of texts P_1, P_2,…, P_n, a measure of smoothness is provided by the sum of the stylistic differences between consecutive texts. If the difference between P_{i+1} and P_i is represented by the symbol $||P_i - P_{i+1}||$, then this sum is $||P_1 - P_2|| + ||P_2 - P_3|| + \ldots + ||P_{n-1} - P_n||$, or in more compact notation:

$$J = \sum_{i=1}^{n-1} ||P_i - P_{i+1}||$$

Now suppose we change the order of the texts, thus reassigning the labels P_i to the texts in a different way. The new permutation may yield a new value for J since the value of J depends on the ordering. A permutation of texts that

reduces J increases smoothness. Smoothness is thus a relative concept, but one can define an absolute version. If J is "significantly" lower than the value of J averaged over all possible permutations, then we have a smooth trajectory. Of course, one may formalize "significantly" in terms of confidence bands.

The quantity $||P_i - P_{i+1}||$ is most straightforward to calculate for univariate data, as P_i is represented by a number (say, average verse length or the frequency of one morpheme). In that case, $||P_i - P_{i+1}||$ is simply the magnitude of the difference between the two numbers. Graphically, it's the difference in height of two consecutive columns. Thus J is the sum of the differences in heights of consecutive columns. J is minimized, and smoothness is maximized, if the columns are arranged in decreasing order of height, or in the reverse order of that. However, there are many smooth trajectories that fall short of maximal smoothness.

In the multivariate case, P_i are represented as vectors (rows) of morpheme frequencies. Arrange P_i as rows on top of one another. One gets a matrix in which each row represents a text (there being n texts) and each column a morpheme (out of m morphemes). At the i'th row and j'th column, we have F_{ij}, the relative frequency count of the j'th morpheme in the i'th text, P_i. The city-block distance between consecutive texts P_i and P_{i+1} is obtained by summing the differences of their morpheme frequencies:

$$||P_i - P_{i+1}|| = \sum_{j=1}^{m} |F_{ij} - F_{(i+1)j}|$$

As before, a measure of smoothness is the sum of such distances between consecutive texts:

$$\mathrm{J} = \sum_{i=1}^{n-1} ||P_i - P_{i+1}|| = \sum_{i=1}^{n-1} \sum_{j=1}^{m} |F_{ij} - F_{(i+1)j}|$$

As in the univariate case, smoothness is inversely related to J.

Incidentally, to reduce or filter out noise or reduce the effects of outlier texts, one may sum a moving average, for example

$$\mathrm{J} = \sum ||P_i - P_{i+1}|| + \alpha ||P_i - P_{i+2}|| + \beta ||P_i - P_{i+3}||$$

where α and β are constants between 0 and 1, and where the summand is defined appropriately for the boundary cases. Obviously, variations on the summand are possible.