

NEH Application Cover Sheet

Digital Humanities Start-up Grants

PROJECT DIRECTOR

Dr. Joshua Berman
Professor of Jewish Studies
c/o Assistant to the President
7605 Old York Road
Melrose Park, PA 19207-3010
UNITED STATES

E-mail: jberman@gratz.edu
Phone(W): 215-635-7300 ext. 133
Phone(H):
Fax: 215-635-1046

Field of Expertise: Literature: Ancient Literature

INSTITUTION

Gratz College
Melrose Park, PA UNITED STATES

APPLICATION INFORMATION

Title: *Scriptorium: A Web Application for Text Categorization and Analysis of the Hebrew Scriptures*

Grant Period: From 5/2014 to 10/2015

Field of Project: Literature: Ancient Literature

Description of Project: Scholars have struggled for more than two centuries to parse the texts of the Hebrew Scriptures into their constituent parts and to identify the different authors and editorial layers of the text. Many have tried to parse the texts on the basis of different styles. The quest, though, for any quantifiable, controllable measure of what "style" is remains elusive. Over the last decade, a handful of computational linguists have achieved unprecedented results working in the subfield of authorship attribution. Scriptorium marshals these advances to produce a web-based application for the categorization and analysis of Hebrew Scriptures along lexical and morphographic lines. Scriptorium will enable scholars to group any combination of texts from within the Hebrew Bible and conduct a range of machine learning based experiments. Scriptorium will exponentially enhance scholars' capacity to classify and categorize groups of texts by style in a controlled and quantifiable manner.

BUDGET

Outright Request	\$59,976.00	Cost Sharing	\$11,160.00
Matching Request		Total Budget	\$71,136.00
Total NEH	\$59,976.00		

GRANT ADMINISTRATOR

Ms. Dodi Klimoff
Administrative Assistant to the President
7605 Old York Road
Melrose Park, PA 19027-3010
UNITED STATES

E-mail: dklimoff@gratz.edu
Phone(W): 215-635-7300 ext. 133
Fax: 215-635-1046

Scriptorium: A Web Application for Text Categorization and Analysis of the Hebrew Scriptures

Table of Contents

Table of Contents	1
Participants	2
Abstract	3
Statement of Innovation	3
Statement of Significance to the Humanities	3
Narrative	4
Enhancing the Humanities Through Innovation	4
Environmental Scan	6
History and Duration of Project	7
Work Plan	8
Staff	9
Project Product and Dissemination	9
Budget	10
Biographies	12
Data Management Plan	13
Letters of Support	14
Appendix A: Letters of Commitment from the Advisory Board	16
Appendix B: Technical Description of the Core Algorithm	18
Appendix C: Illustrations of the User Interface	19
Appendix D: Relevant Publications of Moshe Koppel	22

List of Participants

Berman, Joshua (project director) - Gratz College and Bar-Ilan University

Koppel, Moshe (co-project director) - Bar-Ilan University

Advisory Board

Eli Alshech – Independent Consultant

Maurice, Lisa – Bar-Ilan University

3. Abstract

Scholars have struggled for more than two centuries to parse the texts of the Hebrew Scriptures and to identify the different authors and editorial layers of the text. Many have tried to parse the texts on the basis of different styles. The quest, though, for any quantifiable, controllable measure of what "style" is remains elusive. Over the last decade, a handful of computational linguists have achieved unprecedented results working in the sub-field of authorship attribution. Scriptorium marshals these advances to produce a web-based application for the categorization and analysis of Hebrew Scriptures along lexical and morphographic lines. Scriptorium will enable scholars to input well-attested examples of texts in each of a set of categories and to obtain classification of any disputed text to one of the chosen categories. Scriptorium will exponentially enhance scholars' capacity to classify and categorize groups of texts by style in a controlled and quantifiable manner.

Statement of Innovation

Scriptorium will be the first tool available that avails scholars of pre-modern texts of the advances in authorship attribution algorithms, in their original languages. Relative to other authorship attribution applications now available, Scriptorium will offer a more intuitive user experience and a more precise analytical capacity, particularly concerning the integration of text features, such as morphological analysis, specifically tailored to biblical texts.

Statement of Significance to the Humanities

Scriptorium will give biblical scholars new tools with which to consider a range of critical questions: How many authors may have produced this text? Are two disparate texts written by the same author? Does this text reflect an earlier stage of ancient Hebrew or a later one? Significantly for the humanities more broadly, Scriptorium can easily be adapted to serve as a tool for text categorization and analysis of a range of pre-modern texts, including the entire corpus of Greek and Latin literature.

4. Narrative

Enhancing the Humanities Through Innovation

Categorizing Pre-modern Texts on the Basis of Style: The Current State of Affairs

Scholars across a wide range of disciplines centered around pre-modern texts – biblical studies, Assyriology, classics, medieval studies -- routinely work with texts whose provenance and date cannot be definitively established. A set of standard questions emerge in all of these disciplines about many of the texts under study: is the text in its received form a composite work of more than author? If so, can we determine how many authorial strands it contains? When a text lacks attribution, on what basis may we legitimately attribute it to a known author? If we cannot attribute it to a specific author, can we assess its linguistic register and through that determine the region of its provenance or its approximate era?

While scholars often make determinations to such questions on the basis of the content of a given text, this is not always possible. In many instances scholars attempt to categorize texts and answer such questions on the basis of a text's linguistic characteristics, or, in other words, its style. Telltale markers sometimes stand out and provide clues, and every discipline has its canon of accepted markers of distinctive style through which it categorizes its texts, independent of clues from content.

The methodology employed is typically impressionistic: certain conspicuous linguistic elements trend in highly consistent and distinct fashions. The impressionistic nature of much of this work, like much of the work in literary analysis in general, demands of scholars a great deal of caution regarding possible conclusions that can be made about the text at hand. We propose to provide scholars with an easily-employed tool that marshals the full panoply of measurable textual features and state-of-the-art machine-learning methods to provide reliable statistical evidence confirming (or falsifying) user-provided hypotheses about distinct stylistic threads in the entire biblical corpus.

Categorizing Pre-modern Texts on the Basis of Style: The Contribution of Scriptorium

The field of authorship attribution focuses on the trove of more subtle linguistic characteristics that texts contain. For example, given any biblical text (verse, chapter or book), Scriptorium computes the frequency in the text of each of a huge array of textual features, including individual words, phrases, parts of speech, morphological structures and many more. Given examples of such texts divided into user-chosen categories, machine-learning methods are used to establish patterns of usage characteristic of each category (See Appendix B – Technical Description of the Core Algorithm). For example, Scriptorium's algorithms might find, inter alia, that biblical texts from the post-exilic period use one word for *contract*, while known texts from before the destruction of Jerusalem in 586 b.c.e. use a different word. We know that Shakespeare, for example would employ *thee* and *you* interchangeably as he would *thy* and *your*. His language, as language is at all times, was in a state of flux. Biblical Hebrew, similarly, exhibits two forms for the pronoun *I*. Neither form is exclusive to earlier texts or to later ones, though there is a prevalence for one in pre-exilic Hebrew and for the other in post-exilic Hebrew. Scriptorium calculates a weight for each in determining whether a text is early or late. Many hundreds of similar such subtle

patterns are formally combined to extrapolate a single optimal rule that distinguishes post-exilic from pre-exilic texts. This rule can now be used to classify texts of disputed or indeterminate provenance as post-exilic or pre-exilic. Of course, the same methodology can be used to classify texts according to any division specified by a user, so long as the user can provide a set of texts to establish a rule.

A recurring concern about new tools in the digital humanities is that they will be utilized only by a small number of scholars, often those associated personally or institutionally with the tool's development. In the case of Scriptorium, however, both the range of its capabilities and its ease of use will ensure that all scholars of Hebrew Scriptures will be eager to harness its power. Because its data can and will be applied to so many of the critical questions facing scholars of the Hebrew Bible, we are hopeful that Scriptorium will be recognized as the most significant technological advance in biblical scholarship since the first text-mining tools emerged more than 40 years ago.

Authorship attribution algorithms can often provide outstanding results on the basis of lexicography alone. For practical application, therefore, a digitized text is all that is needed. However, sometimes results can be best achieved by analyzing a text's morphological characteristics as well. In these cases, it is insufficient to provide the algorithm a digitized text; the text must be morphologically annotated as well. The Hebrew Bible serves as a good test corpus for a tool of this kind, as it is already fully annotated morphologically in digitized form in several commercially available software applications: part-of-speech tagging, syntactic parsing, and dependency parsing to decompose sentences into their grammatical constituents. In fact, the availability of canonical translations of the entire biblical corpus permits us to automatically identify synonym sets, which are especially useful as features for authorship attribution.

To summarize, the tool that we propose to put in the hands of scholars will:

- include an interface especially designed to allow easy specification of any biblical texts (See Appendix C – Illustrations of User Interface).
- come pre-loaded with the entire Hebrew Bible, including all lexical, morphological and syntactical data regarding each word, verse, chapter and book. The original Hebrew texts will be taken from Bar-Ilan University's Responsa project¹ and tags will be based, inter alia, on Westminster Hebrew morphology. (While our software will be made freely available, the texts themselves will be made available in accordance with licenses issued by data providers.)
- display automatically learned classifiers that distinguish between any user-provided categories of texts
- classify all biblical texts in one of the user-provided categories, display the main markers in the text that explain its classification and provide the degree of certainty of that classification

Thus, biblical scholars will be able to exploit all available measurable information regarding a text in order to verify or falsify hypotheses about apparent divisions of the biblical oeuvre into distinct stylistic threads.

¹ <http://www.biu.ac.il/jh/Responsa/>

We note further that the entirety of Greek and Latin literature have been annotated in similar fashion as well in databases such as the Perseus Digital Library,² The Library of Latin Texts,³ which covers Latin Literature until Vatican II (1964) and the Thesaurus Linguae Graecae⁴ – which covers Greek literature from Homer until the 15th c. Working in conjunction with a scholar of Islamic studies, Eli Alshech, Koppel has already demonstrated that the algorithm is an effective tool for the categorization of texts in Arabic (See advisory board letter of commitment by Eli Alshech).⁵ The algorithm that will be underlying Scriptorium can subsequently be adapted to create a similar tool for the study of Greek and Latin literature as well.

Environmental Scan

In the past several decades scholars of the Hebrew Bible have attempted to delineate the style of a given book of the Bible based on impressionistic methods. Yet, by and large the attempts to delineate a thoroughgoing analysis of a text's distinct style have been rejected out of concern for statistical and methodological flaws and the lack of sufficient control.⁶

Powerful text-mining applications such as Logos and Accordance allow for sophisticated lexicographic and morphographic searches of Hebrew Scripture.⁷ Accordance provides statistical analysis for the distribution of any given word or morphological marker. But none of these use cutting-edge methods to weight the constellation of thousands of markers together toward the end of delineating a text class or category.

These same limitations characterize the tools for textual analysis available within the humanities generally today. Some applications supply morphological annotation to a text, such as WordHoard and Stanford CoreNLP.⁸ The tools indexed in the Text Analysis Portal for Research (TAPoR) and in the MONK (Metadata Offer New Knowledge) project provide a range of services that enhance display, mining and cross-reference of texts.⁹ None execute the simultaneous analysis of hundreds of linguistic markers to determine text categorization.

We are encouraged that ODH has recently funded a start-up grant for a project employing the algorithm of another scholar in this field, Patrick Juola, in project # HD-51556-12, "Is That You, Mr. Lincoln: Applying Authorship Attribution to the Early Political Writings of Abraham Lincoln."

Scriptorium differs in two ways. First, Scriptorium is especially designed for working with pre-modern texts, specifically, in our work with the Hebrew Bible. Second, our

² <http://www.perseus.tufts.edu/hopper/>

³ http://www.brepols.net/publishers/pdf/Brepolis_LLT_En.pdf

⁴ <http://www.tlg.uci.edu/>

⁵ <http://u.cs.biu.ac.il/~koppel/papers/arabictextcat-isi-final-koppel.pdf>

⁶ See in recent discussion, Richard A. O'Keefe, "Critical Remarks on Houk's 'Statistical Analysis of Genesis Sources'," *Journal for the Study of the Old Testament*, 29:4 (2005) 409-437.

⁷ <http://www.logos.com/>; <http://www.accordancebible.com/>.

⁸ <http://wordhoard.northwestern.edu/userman/index.html>;
<http://nlp.stanford.edu/software/corenlp.shtml>.

⁹ <http://www.tapor.ca/>; <http://www.monkproject.org/>.

project does not execute an experiment on a limited set of texts, important as they are. Rather, it creates a tool for scholars to conduct their own experiments on a vastly wider corpus of literature.

Some recent software packages used in the authorship attribution community incorporate a number of the elements we will include in our system. These include Juola's JGAAP package,¹⁰ JStylo,¹¹ Signature,¹² Writeprint Stylometry,¹³ and the Delta Spreadsheets.¹⁴ Unlike these systems, however, ours will be especially designed to exploit precisely the textual characteristics such as morphology that are useful for pre-modern literary texts such as the Bible and will include a graphical user interface especially designed for use by humanities scholars working with pre-modern literary texts in classical languages.

Scriptorium will incorporate machine learning tools of the most sophisticated sort alongside an intuitive interface especially designed for scholars of pre-modern texts, especially the Bible. The sophisticated user will be able to choose from a vast array of feature sets and learning algorithms. However, the scholar who does not wish to grapple with the intricacies of advanced software will be able to take advantage of default settings optimally chosen on the basis of the developers' expertise in handling these sorts of problems.

In addition, the user will be spared the need to upload data and to deal with pre-processing the data. The entire biblical corpus, including all encoding, tagging and cleaning, will be pre-loaded, leaving the user only to indicate (from a drop-down menu) which texts to use as training examples for each class under consideration. The pre-loaded data will leverage all the manual tagging available for biblical literature (translations, synonyms, morphology, disambiguation of homonyms, etc.).

The output that we will provide will also be richer than that currently available. In particular, we will display the features that best distinguish between two or more training sets and we will also show which of these are responsible for a particular determination regarding a given test document.

History and Duration of the Project – This project is based upon ten years of research by Koppel in the area of authorship attribution. (See Appendix D for a list of relevant papers). Koppel's work includes novel methods for determining authorship of an anonymous document from among a small closed set of candidates¹⁵, methods for providing a demographic profile of an anonymous author¹⁶ and, more recently, determining authorship from among a very large (thousands of authors) open set of

¹⁰ http://evllabs.com/jgaap/w/index.php/Main_Page

¹¹ https://psal.cs.drexel.edu/index.php/Main_Page

¹² <http://www.philocomp.net/?pageref=humanities&page=signature>

¹³ http://www.downloadplex.com/Scripts/PHP/Modules/writeprint-stylometry-scripts_330446.html

¹⁴ <https://files.nyu.edu/dh3/public/TheDeltaSpreadsheets.html>

¹⁵ Koppel, M., Schler, J. and Argamon, S. (2009), [Computational Methods in Authorship Attribution](#), JASIST, 60 (1): 9-26

¹⁶ S. Argamon, M. Koppel, J. Pennebaker and J. Schler (2009), [Automatically Profiling the Author of an Anonymous Text](#), Communications of the ACM, 52 (2): 119-123 (virtual extension).

candidates¹⁷. All this work was accompanied by experimental results performed on proprietary software.

In the past few years, Koppel has considered the problem of decomposing a multi-author document into authorial threads.¹⁸ In experiments on the Hebrew Bible, these methods proved startlingly precise and have garnered much attention.¹⁹ Koppel took all the verses of two lengthy biblical books describing the same period of biblical history--Jeremiah and Ezekiel--and effectively shuffled them together, creating a single merged text. Without any other input, he ordered the algorithm to split the file by style into two sets of verses. The algorithm could assign a verse to set A, to set B, or could reserve judgment on a decision. For approximately 19% of the verses, the algorithm made no decision. For the remaining 81%, the algorithm split the verses into two sets, that with 99% accuracy accorded to the books of Jeremiah and Ezekiel. Koppel repeated this experiment with five other pairings of biblical books of similar genre. The results were replicated time and again: the algorithm would assign approximately 80% of the verses to two sets reflecting the books in question, and would do so with extremely high accuracy. The time is now ripe to further develop this software so that it can be usable by scholars of biblical studies, with no expertise in the implemented computational methods. Berman's expertise in biblical studies will facilitate development of a system optimally designed to meet the needs of the scholarly community.

The first phase of the project, which is covered by this proposal, includes specially adapted tools for handling cases where users provide labeled examples of texts in each specified category--such as a hypothesized common author-- and these are exploited to find a classifier based on the statistically significant common characteristics of those texts. This approach is called "supervised" learning. This part of the project should be completed within 12 months of commencement. This stage of the project allows scholars to establish a set of "baseline" texts, against which another text may be determined to belong or not belong to that set. In subsequent stages of the initial project, we will enable scholars to separate texts into authorial threads, even in the absence of user-provided sets of "baseline" texts, or what is called "unsupervised" learning.)

Work Plan - In the first phase of this Level II project, Berman will lead the effort to catalog the range of text classification questions that biblicists routinely raise, so as to define the features that will make for an effective application. (May-June 2014). In the second phase, Koppel and the developers will design natural language processing tools to automatically extract all these features from a text (July-December 2014). In the next phase, Koppel and Berman will instruct the developers on how to adapt our text categorization tools to the problems at hand, focusing especially on developing a graphical user interface that meets the needs of scholars not expert in our methods (January-June 2015). In the final phase Berman will coordinate the effort to test the

¹⁷ Koppel, M. Schler, J. and S. Argamon (2011), [Authorship Attribution in the Wild](#), *Language Resources and Evaluation* 45(1) (special issue on Plagiarism and Authorship Analysis)

¹⁸ M. Koppel, N. Akiva, I. Dershowitz and N. Dershowitz, (2011). [Unsupervised Decomposition of a Document Into Authorial Components](#), *Proceedings of ACL*, Portland OR, June 2011, pp. 1356-1364.

¹⁹ <http://blogs.wsj.com/tech-europe/tag/moshe-koppel/>; <http://news.discovery.com/tech/bible-algorithm-authors-111013.html>.

efficacy of our methods on the central open problems in the field of biblical studies and adapt the system as necessary. Berman will consult with colleagues at Bar-Ilan University, which hosts the world's largest department of Hebrew Bible Studies (July-October 2014). Berman will consult with advisory board member Lisa Maurice concerning adaptation of the tool for use with texts in Greek and Latin. Eli Alshech will advise on how the tool may be adapted for Arabic language texts. We note that all coding will be done in Java.

Staff

Joshua Berman will be responsible for identifying the scope of questions raised by biblicalists when seeking to classify biblical texts. He will assist Koppel in translating those questions into application features and in developing an effective user interface. He will also be responsible for application documentation, and the overall administration of the project. He will devote 15% of his time to this project, with 3% of his salary to come from the grant budget, and 12% to be cost-shared.

Moshe Koppel will be responsible for developing all algorithms and adapting them to the biblical context. He will oversee all software development and quality assurance. He will devote 15% of his time to the project, with his salary being cost-shared by Bar-Ilan University.

Software Developers - In addition, the staff will include two programmers with the following qualifications:

- expertise in Java
- background in natural language processing,
- experience in interface design
- familiarity with biblical Hebrew

Start-up project product and dissemination

During the course of the grant-funded period, we will complete a working web-application that will be freely available via a dedicated public website. Scriptorium will be tested by faculty members of the Bar-Ilan University Department of Hebrew Bible Studies and will incorporate design refinements based on their feedback. Once the site is public, scholars will be able to test the application and provide additional feedback. The project director and co-director will co-author a white paper. The white paper will focus on the steps that will be necessary to adapt Scriptorium for the analysis of text corpora in Greek, Latin and Arabic. The project director and co-director will present the project at a number of conferences in their respective fields, such as the Society for Biblical Literature and the Association of Computational Linguistics.

We hope to obtain feedback from scholars on the first stage of the project (the one for which we now seek funding) to improve our system. Ongoing development will allow us to more deeply analyze syntactic tagging as well. We intend to subsequently apply for an implementation grant that will allow us to incorporate these improvements as well as to extend these tools to classical literature generally.



Budget Form

Applicant Institution: *Gratz College*
Project Director: *Joshua Berman*
Project Grant Period: *05/01/14-10/31/15*

[click for Budget Instructions](#)

	Computational Details/Notes	(notes)	Year 1	(notes)	Year 2	(notes)	Year 3	Project Total
			05/01/14- 12/31/14		01/01/15-10/31/15		01/01/20__- 12/31/20__	
1. Salaries & Wages								
Project director (Joshua Berman)	Total yearly salary: \$72000	3%	\$2,160	4%	\$2,880	%		\$5,040
Software Developer 1 (TBD)	Annual Salary: 74,000	12%	\$8,880	13%	\$9,620	%		\$18,500
Software Developer 2 (TBD)	annual Salary: 74,000	12%	\$8,880	13%	\$9,620	%		\$18,500
		%		%		%		\$0
		%		%		%		\$0
		%		%		%		\$0
2. Fringe Benefits								
Project Director		16%	\$346		\$461			\$807
Developer 1		16%	\$1,421		\$1,539			\$2,960
Developer 2		16%	\$1,421		\$1,539			
3. Consultant Fees								
								\$0
4. Travel								
Project Director	Travel to NEH Meeting		\$800					\$0
								\$0
5. Supplies & Materials								
								\$0
6. Services								
								\$0
7. Other Costs								
								\$0
8. Total Direct Costs	Per Year		\$23,908		\$25,659		\$0	\$49,567
9. Total Indirect Costs								

Estimated rate: 21% to be negotiated	Per Year	21%	\$5,021		\$5,388		\$0	\$10,409
10. Total Project Costs	(Direct and Indirect costs for entire project)							\$59,976
11. Project Funding								
	a. Requested from NEH				Outright:		\$59,976	
					Federal Matching Funds:		\$0	
					TOTAL REQUESTED FROM NEH:		\$59,976	
	b. Cost Sharing				Applicant's Contributions:		\$0	
	Salary from Gratz and Bar-Ilan for Project Director and Co-director				Third-Party Contributions:		\$11,160	
					Project Income:		\$0	
					Other Federal Agencies:		\$0	
					TOTAL COST SHARING:		\$11,160	
12. Total Project Funding								\$71,136

Total Project Costs must be equal to Total Project Funding ----> (\$59,976 = \$71,136 ?)
 Third-Party Contributions must be

6. Participant Biographies

Joshua Berman has been a professor of Hebrew Bible at Gratz College since 2004 and at Bar-Ilan University since 2002. He focused on aspects of comparing and contrasting biblical texts in his *Narrative Analogy in the Hebrew Bible: Battle Stories and Their Equivalent Non-battle Narratives* (Leiden: Brill, 2004). He is also the author of *Created Equal: How the Bible Broke with Ancient Political Thought* (New York: Oxford University Press, 2008), which was a National Jewish Book Award Finalist in Scholarship in 2008. The primary focus of his work has been the close reading of biblical texts, and his articles have appeared in the major journals of biblical studies, including the *Journal of Biblical Literature*, *Catholic Biblical Quarterly*, *Zeitschrift fue die Alttestamentliche Wissenschaft*, and *Vetus Testamentum*.

Moshe Koppel is a Professor of Computer Science at Bar-Ilan University. He received his PhD from Courant Institute and did post-doctoral work in the Institute for Advanced Study in Princeton. Koppel's research has focused mainly on text categorization, especially authorship attribution. Koppel and his team have developed profiling techniques that successfully determine an author's gender, age, native language, and personality type, using nothing but statistical properties of the author's written work and have also made important advances on authorship problems in which the candidate set might include tens of thousands of candidates. He has also developed techniques for decomposing multi-author documents into authorial threads, with applications to biblical works. Koppel also publishes on social choice theory and has served as an adviser on constitutional affairs in Israel's Knesset. [See Appendix D for Koppel's relevant publications]

Advisory Board

Eli Alshech is currently an independent consultant to the Israeli government on global jihad and Islamic Law, and is a former director at the Middle East Media Research Institute (MEMRI). He holds a PhD in Near Eastern Studies from Princeton University and has taught Islamic law and Islamic history at UCLA. His articles on Islamic law have appeared in *Islamic Law and Society*, *Die Welt des Islams*, *Journal of Near Eastern Studies* and *International Journal of Middle East Studies*. Together with Moshe Koppel he has coauthored a study using the algorithms at the core of the proposed tool for the classification of Arabic language texts. Their work appeared in *In Proceedings of The IEEE Intelligence and Security Informatics (ISI)*, Dallas, Texas, August 2009.

Lisa Maurice is a lecturer in Greek and Latin literature at Bar-Ilan University. She is the author of *The Teacher in Ancient Rome: the Magister and his World* (Lexington, 2013), as well as a number of articles on Roman comedy published in a range of prestigious journals, such as *Mnemosyne*, *Scholia*, and *Syllecta Classica*. Her work over the last few years has been in the field of classical reception and modern popular culture, in which she has published widely, and recently organised an international conference on that subject. In particular she has been working on the reception of classics in children's literature, and is editing a volume on this topic for the Brill's *Metaforms* series, forthcoming in 2014.

7. Data Management Plan

The data to be generated consists of:

1. biblical texts tagged in a variety of ways (lexical, morphological and other), all uniformly formatted for use with our software
2. software for classifying biblical texts in user-provided categories

The algorithm underlying Scriptorium is used in a number of applications in the fields of industry and security and cannot be made available in open source code. Moshe Koppel affirms that those parts of his proprietary software that are necessary for the functioning of the proposed project will be made freely available to the project in perpetuity. Moshe Koppel is committed to working with scholars interested in further development of the proposed tool to accommodate other text corpora in other languages. The field of authorship attribution has a keen interest in the practical application of its findings and such future development represents a core scholarly interest.

Our software will be free for public access (for research purposes). The code will be developed in Java and use published and state of the art algorithms for authorship attribution.

Concerning the housing of the code and the algorithm, we will provide both (a) a stand-alone solution that can be run on end-user's machine and (b) a web based interface to process and analyze the data. The actual code, integration and build of the system will be done in back office, and hosted on web based source control. Data and source code will be stored on web based configuration management services (GitHub).

The application will be offered both (a) as a web based service (no download required) and (b) as a downloadable application. We will make the formatted tagged texts available both online (and will be used as part of the service) as well as integrated within the application and used from there. Users will be able to download the executable code, as well as access the service online (in a service mode - without need to download and install an executable).

All products of research done using our software and data will be the property of the researcher using the program. It is expected that researchers reporting findings attained through Scriptorium will retain their data and share with the wider community as requested, as with the data of any reported scientific experiment.



August 20, 2012

Brett Bobley
Director/CIO
Office of Digital Humanities
National Endowment for the Humanities
Dear Brett,

Moshe Koppel has asked for a letter of support for a Digital Humanities Start-Up Grant for this project entitled *Scriptorium: A Web Portal for Text Categorization and Analysis of the Hebrew Scriptures*. As you probably know already, text categorization and authorship analysis are key problems in the humanities, and being able to do them “digitally” is an important new approach. I’ve been working in the area myself for over a decade and know how challenging it is.

With that said, the Hebrew scriptures in particular are challenging – I’d not touch them myself with a hay-fork for precisely that reason. Not only are they extremely important and well-studied (more people care about Scriptural authorship than possibly any other set of documents in history) but they have been copied, recopied, edited, and generally messed with so many times that the authorial and editorial layers are buried as deeply as an onion. Controversies run rife among traditional scholars and adding computers to the mix is likely to create as many new discussions as it solves.

Moshe is probably the only person in the world who can do it. He’s a top-flight and world-renowned authorship attribution specialist, but also a Talmud and scriptural expert. He’s one of the few people in the field familiar enough with Scriptural Hebrew to address the quirks of that particular language/dialect, and he’s worked specifically with Biblical authorship before. But the true stroke of brilliance behind this project is the way he’s crowdsourcing his own technology via a portal to leverage not only his own knowledge but the “crowd” of Scriptural experts who are not themselves computer-savvy but who can support and develop new methods, features, and experiments by bringing their own traditional scholarship to bear.

Obviously, I support this project – I think that it’s a unique project that could only be done by one group in the world, and that it addresses one of the most important authorship questions. I recommend it for funding.

Patrick Juola
EVL Lab
Duquesne University

Professor Steven Grosby
August 14, 2013

Office of Digital Humanities
National Endowment of the Humanities

Re: Grant Application "Scriptorium: A Web Application for Text Categorization and Analysis of the Hebrew Scriptures"

I write in support of the application "Scriptorium." I do so as a Professor of Religion and biblical scholar because of the probable, significant usefulness that the proposed web-based application would likely have on the humanities and my field of study in particular. It would do so by providing an objective means of separating out apparently distinct authorial strains of writings from literary works that have long been thought to be composites.

As is well known, for the last four to five hundred years, many biblical scholars and other humanists have raised important questions about both the compositional unity of the Pentateuch (the so-called "documentary hypothesis") and an underlying document of the New Testament (the existence of "Q", as a compilation of the sayings of Jesus, or variations of such a possibility as the Gospel of Mark being the basis for Matthew and Luke). Actually, these questions were raised earlier in the middle ages and even in antiquity. Despite often brilliant analyses in response to the questions, no agreement exists. Although the details of the algorithms of technology involved, as described in Moshe Koppel et al, "Unsupervised Decomposition of a Document into Authorial Components," are beyond my understanding, it is very easy for me to imagine that a user-friendly web application of that technology would likely be an exciting contribution to these centuries-long scholarly preoccupations.

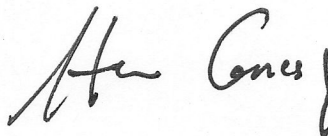
If I understand properly the outlines of the technology, it would do so by providing impartial, that is, objective means for grouping together, and thereby isolating, apparent authorial strains in contrast to the criteria used by the analyst. Certain similarities, mostly based on vocabulary, have long been observed. For example, centuries ago, it was observed that in some chapters of Genesis the deity is referred to as *elohim*, while in other chapters—in the tradition of the term's designation—*adonai*; and then in yet other chapters, a combination of both. Does this variation indicate different authors of respectively different compositional traditions that have been subsequently brought together? To take another example, Exodus 23:10-11, Deuteronomy 24:19-22, and Leviticus 19:9-10 deal with the same legal problem; but the vocabulary employed differs among these three formulations. Here again, do we have different authors and even possibly different historical and/or institutional settings for what appear to be different literary groupings? If there are, in fact, different authorial components, are there other ways to group them that would confirm or disconfirm the apparent groupings based upon previous scholarly



assumptions? It seems that what is being proposed to the NEH is precisely a different way. Now, if so, the resulting patterns will themselves be subject to study. Having them isolated, the analyst can then see what patterns emerge, rather than, in an all too often circular fashion, finding a pattern confirmed when the pattern is already stipulated.

Evidently, earlier applications of the program have confirmed its merit, for example, when Koppel et al grouped together blindly the verses from Jeremiah and Ezekiel, and then with remarkable accuracy the program with its algorithms succeeded in distinguishing the two books. However, other interesting possibilities of the application of this program may point beyond authorial segregation, for it may also be that one might also be able to draw other conclusions about the emergent authorial strains, for example, their setting and even dating relative to other strains. For example, we assume that Joshua 10:12-13 and 2 Samuel 1:19-27 are relatively old because they are designated as "songs" from the "Book of Jashar" (other presumably older material would be "The Song of Moses" in Exodus 15 and the "Song of Deborah" in Judges 5.) The assumption is based on their designation as songs (hence, the probable products of an older oral tradition), their description as being from the lost "Book of Jashar," and occasional lexical and morphological features of what may be an archaic Hebrew. One wonders, however, what would be the results of running the verses through this program? Would a different, wider strain of similar vocabulary and morphological features emerge? And if so, what ought the scholar to conclude?

In conclusion, what is being proposed seems to me to represent not only a new technological and digital tool for the humanities, but also one with exciting possibilities that would quickly be realized by scholars. At a minimum, many long held assumptions would be impartially confirmed (or disconfirmed, or at least no longer assumed to be so). But beyond that minimum, other provocatively new lines of inquiry might be opened up, even if, as with the possible dating of these authorial strains, these would have to be pursued with appropriate scholarly caution. If, as I suspect, new lines of inquiry were opened up, then this tool might have important bearing well beyond its initial literary application; for it would indicate implications for our historical understanding, including areas of legal and institutional development. Finally, given my own area of scholarly expertise, I can, with complete confidence based upon my knowledge of his work, judge Professor Berman as being well qualified to serve as the project director.



Steven Grosby
Professor of Religion

Bar-Ilan University (RA)
52900 Ramat-Gan, Israel
Department of Classical Studies



Bar-Ilan University

Tel: 03-5318231 :טל
classics@biu.ac.il
Fax: 972-3-5353937 :פקס

Email: classics@biu.ac.il

בס"ד
אוניברסיטת בר-אילן (ע"ר)
רמת-גן 52900
המחלקה ללימודים קלאסיים

27th August 2013

NEH Digital Humanities Start-Up Grant

I am delighted to confirm my willingness to serve on the advisory board for the project "Scriptorium: A Web Application for Text Categorization and Analysis of the Hebrew Scriptures," to be conducted by Dr. Joshua Berman and Dr. Moshe Koppel. There is a considerable overlap in interest between scholars of the Hebrew Bible and those of classical studies since both groups typically are concerned with texts, the authorship, influence and chronology of which are uncertain. Indeed, textual criticism in its modern sense which is so central to classical studies itself developed originally as a discipline concerned with the Bible. As the *apparatus criticus* of any classical text reflects, the question of what the ancient author actually wrote is one that has long been debated by scholars and which remains fundamental to classical studies. The incomplete nature of texts, the manuscript tradition, the original oral composition of some texts that were only written down at a later stage, and the very fact that even the earliest manuscripts of texts in the field of classical studies were written about a millennium after their original composition, highlight the difficulties faced by scholars attempting to examine texts, whether canonical (texts of Virgil or Livy) or not (papyri, fragments etc.). That these texts were often based on other, earlier writings makes these questions even more fascinating - the whole subspecialisation of source criticism is in fact dedicated to trying to determine the earlier sources used by the classical authors. Similarly, the debates about the provenance of individual words, lines and even whole works all reflect the centrality of questions of authorship to the study of Latin and Greek texts.

It seems to me extremely likely therefore that, assuming that Hebrew Bible scholars do find this tool helpful, there is reason to believe that such a tool, properly adapted, could be of service for scholars in the field of classical studies as well. I would be very interested, following development of the software, in assisting in the consideration of how to adapt this tool to serve the needs of scholars working in Classics.

With all best wishes,

Dr. Lisa Maurice

08/09/2013

Dear Moshe,

Thank you for your request to serve on the advisory board of your project concerning computerized authorship studies for biblical texts. I will be very happy to do so.

As you know, my expertise lies in classical Arabic Studies, the area in which I did my doctoral work in Princeton University under the guidance of Prof. Michal Cook. Given my familiarity with your work and our long cooperation in this field I am convinced that the tools you are designing would be invaluable in the area of Biblical Studies as well as Quranic Studies. I would be delighted to advise you as to how to extend your work from biblical texts to classical Arabic texts.

I wish you much success.

Dr. Eli Alshech
ealshech@gmail.com



Advisor in the Office of the Prime Minister of Israel
Research Fellow at the Hebrew University

Appendix B – Technical Description of the Core Algorithm

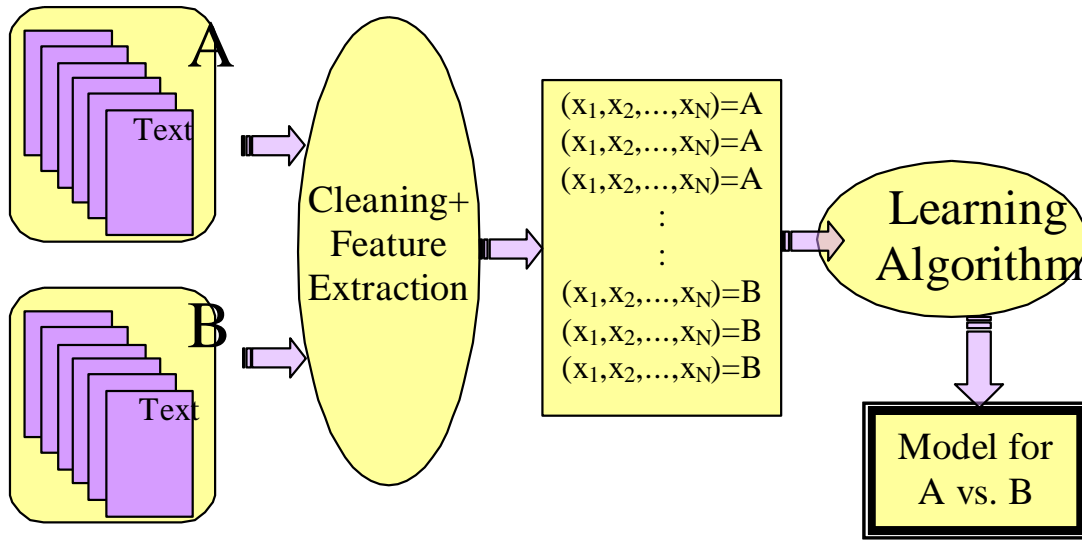


Figure 1: Architecture of a text categorization system.

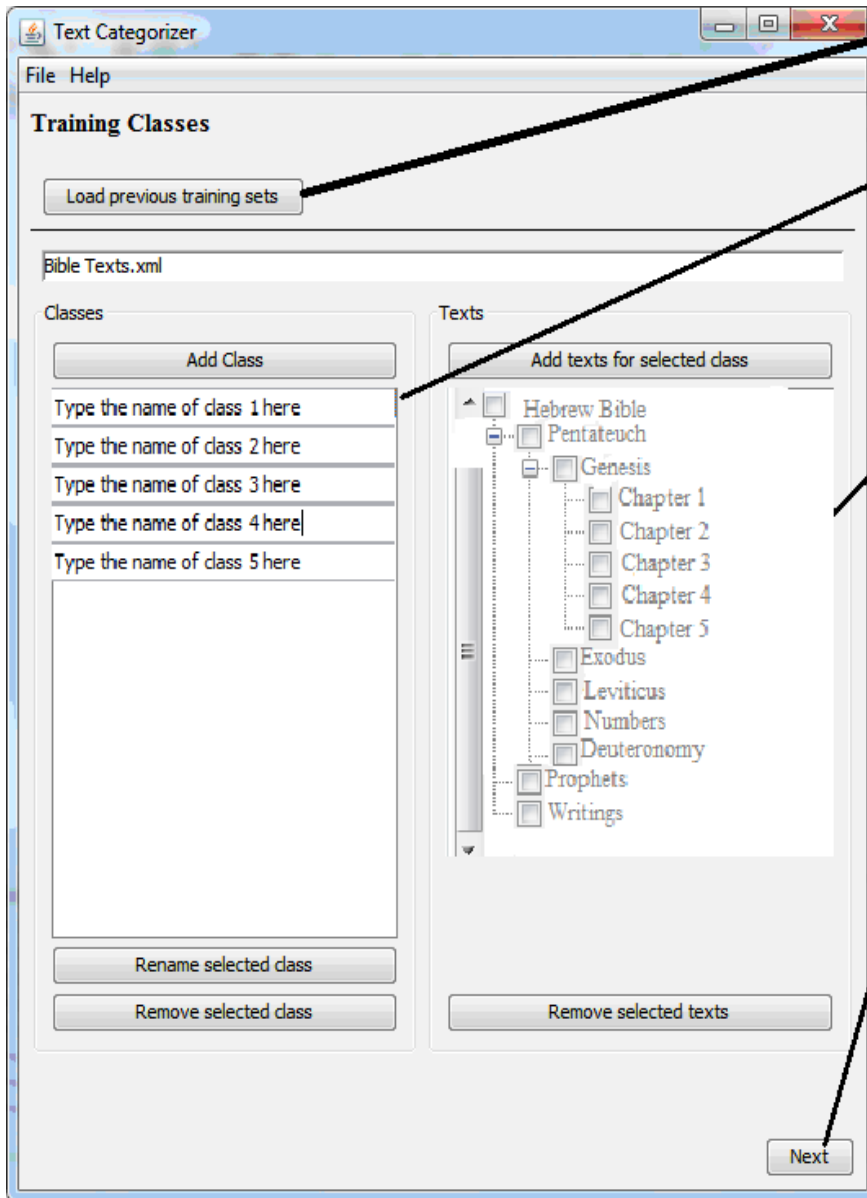
In Figure 1 we show the basic architecture of a text categorization system in which we are given examples of two classes of documents, Class A and Class B. The first step, document representation, involves defining a set of text features which might potentially be useful for categorizing texts in a given corpus and then representing each text as a vector in which entries represent (some non-decreasing function of) the frequency of each feature in the text. Once documents have been represented as vectors, a learning algorithm is used to construct models that distinguish between vectors representing documents in Class A and vectors representing documents in Class B.

We briefly review the feature types and learning methods that will be available for our purposes.

Since we wish to classify biblical texts according to stylistic (rather than topical) considerations, we wish to use linguistic features that are content-independent. In the past researchers have used for this purpose lexical features such as function words, syntactic features, or complexity-based feature such as word and sentence length). We will make all the standard feature types available. In addition, we note that Hebrew texts present special problems in terms of feature selection for style-based classification, since function words tend to be conflated into word affixes in Hebrew, thus decreasing the number of function words but increasing the amount of morphological features that can be exploited. We will make such morphological features available.

As for learning algorithms, our software will be compatible with the WEKA package, so that all the algorithms available in WEKA will be available, including SMO, J4.8, Naïve Bayes, Logistic Regression, kNN, etc.

Appendix C: Illustrations of User Interface



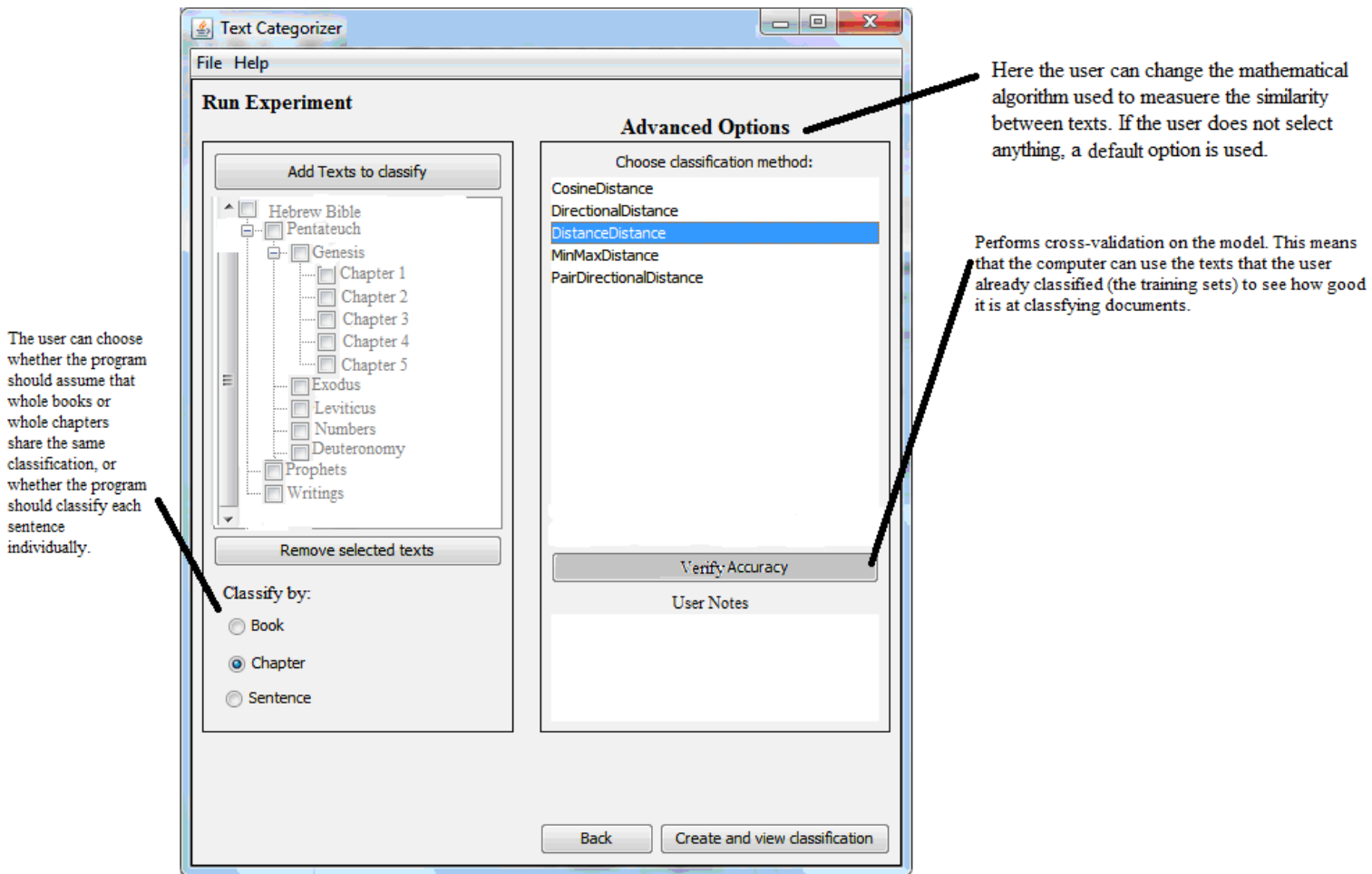
If a user had previously made a training set, they can simply load that and skip the rest of this page.

Users can add any amount of classes, and name them whatever they wish (e.g Old Biblical Hebrew, New Biblical Hebrew, etc.)

Users can add any combination of complete books, chapters, or verses to any of the classes.

A user who wishes to go on to the next page may do so by clicking on this button.

On this screen, a researcher begins to set up her experiment. Let us suppose that she wishes to determine whether the book of Deuteronomy, whose date is debated within the scholarship, reflects the style of books written before the exile to Babylon (Early Hebrew), or the style of books composed following the return to Zion (Late Hebrew). On this screen she will define two training classes. In one, she will load books that define the class Early Hebrew. In her second class, she will load books that define Late Hebrew.



Our researcher chooses the Book of Deuteronomy as the text that she wishes to classify. She can choose the type of results she wishes to receive. She may wish to know, simply, whether the book of Deuteronomy as a whole reflects early or late Hebrew. She may wish to get results for each chapter of that book; perhaps some chapters reflect early Hebrew, while perhaps other chapters were composed and added to the book later. Or, she may wish to examine each and every sentence of the book. It may turn out that the while most of the book reflects early Hebrew, a handful of sentences will reflect late Hebrew and will suggest that these were later additions to the book. In this illustration, she has chosen to analyze the book by chapter.

These values reflect the significance of the feature for classifying a document in a particular class.

This shows which class each document was assigned to. In this case, the program determined which chapters used Early Biblical Hebrew, and which chapters used Late Biblical Hebrew.

This pane shows statistics on all of the features (in this case, the words) that were used to make the classification.

This pane shows one of the biblical chapters which was classified as using Early Biblical Hebrew. Words that helped determine this are highlighted.

This is called a confusion matrix. It shows how well the program categorized the texts. In this case, the program appears to have been perfect.

This button allows the user to confirm that everything is indeed OK.

Ind	Name	Assigned
13	Deuteronomy Chapter 1	EarlyHebrew
14	Deuteronomy Chapter 2	EarlyHebrew
15	Deuteronomy Chapter 3	EarlyHebrew
16	Deuteronomy Chapter 4	EarlyHebrew
17	Deuteronomy Chapter 5	EarlyHebrew
18	Deuteronomy Chapter 6	EarlyHebrew

Ind	Feature	Count	FA Early...	FA LateH...	WDA Earl...
3	יקוק	1759	0.017061	0.010561	1.703389
1	את	2906	0.027485	0.018712	1.126555
2	אשר	2004	0.018652	0.013353	1.057409
12	לא	756	0.007028	0.005006	0.604497
51	היום	211	0.002413	0.000718	0.287210
10	כל	871	0.007419	0.006857	0.127027
56	לי	189	0.001915	0.001004	0.126153
202	אנכי	71	0.000925	0.000063	0.113154
194	ארץ	73	0.000766	0.000352	0.108989
91	אתו	137	0.001608	0.000404	0.098373
82	אסרי	154	0.001774	0.000503	0.067788
535	מים	29	0.000348	0.000064	0.055728
670	פן	24	0.000314	0.000022	0.049456
529	נחשת	30	0.000288	0.000190	0.045272
76	שנה	161	0.001463	0.001140	0.044567
217	והנה	66	0.000828	0.000100	0.044196
101	לך	127	0.001304	0.000649	0.038947
55	עשה	197	0.001986	0.001075	0.037350
405	אלהיה	39	0.000394	0.000211	0.033041
360	מבית	43	0.000438	0.000235	0.032599
791	לבלתי	20	0.000233	0.000063	0.032098
758	לך	21	0.000249	0.000044	0.028897
876	מים	18	0.000204	0.000063	0.018803

Ind	*	EarlyHebrew	LateHebrew
1	EarlyHebrew	406	0
2	LateHebrew	0	294

Color By Features Color By Features Set

Mark Features of Scanner: [Dropdown]

Color By Features Color By Features Set

Mark Top 10 Features Mark Selected Rows

OK

In this frame, the researcher receives her results. As per her selection, she is given the analysis of the Book of Deuteronomy by chapter. Here the results are presented for Deuteronomy chapter 5. The highlighted text at the lower left indicates the lexemes in the chapter that contributed to the determination that this chapter reflects early biblical Hebrew. The table on the upper right presents the terms in order of statistical significance that contributed to this classification.

Appendix B. Koppel's Relevant Publications:

- M.Koppel, S. Argamon and A. Shimoni (2003), [Automatically categorizing written texts by author gender](#), *Literary and Linguistic Computing* 17(4), November 2002, pp. 401-412.
- S. Argamon, M. Koppel, J. Fine and A. Shimoni (2003), [Gender, Genre, and Writing Style in Formal Written Texts](#), *Text*, 23(3), August 2003.
- M. Koppel and J. Schler (2003), [Exploiting Stylistic Idiosyncrasies for Authorship Attribution](#), in *Proceedings of IJCAI'03 Workshop on Computational Approaches to Style Analysis and Synthesis*, Acapulco, Mexico.
- M. Koppel, N. Akiva and I. Dagan (2003), [A Corpus-Independent Feature Set for Style Based Text Categorization](#), in *Proceedings of IJCAI'03 Workshop on Computational Approaches to Style Analysis and Synthesis*, Acapulco, Mexico.
- M. Koppel and J. Schler (2004), [Authorship Verification as a One-Class Classification Problem](#), in *Proceedings of 21st International Conference on Machine Learning*, July 2004, Banff, Canada, pp. 489-495.
- S. Argamon, S. Dawhle, M. Koppel and J. Pennebaker (2005), [Lexical Predictors of Personality Type](#), *Proceedings of Classification Society of North America, St. Louis MI, June 2005*.
- M. Koppel and J. Schler (2005), [Using Neutral Examples for Learning Polarity](#) (poster), *Proceedings of IJCAI, Edinburgh, Scotland, July 2005*.
- M. Koppel, J. Schler and K. Zigdon (2005), [Determining an Author's Native Language by Mining a Text for Errors \(short paper\)](#), *Proceedings of KDD, Chicago IL, August 2005*.
- M. Koppel, D. Mughaz and N. Akiva (2006), [New Methods for Attribution of Rabbinic Literature](#), *Hebrew Linguistics: A Journal for Hebrew Descriptive, Computational and Applied Linguistics*, to appear.
- J. Schler, Moshe Koppel, S. Argamon and J. Pennebaker (2006), [Effects of Age and Gender on Blogging](#), in *Proc. of AAAI Spring Symposium on Computational Approaches for Analyzing Weblogs, March 2006*.
- S. Hota, S. Argamon, M. Koppel, I. Zigdon (2006). [Performing Gender: Automatic Stylistic Analysis of Shakespeare's Characters](#), in *Proc. Digital Humanities, July 2006*.
- M. Koppel, J. Schler, S. Argamon and E. Messeri (2006). [Authorship Attribution with Thousands of Candidate Authors](#), (poster). in *Proc. Of 29th Annual International ACM SIGIR Conference on Research & Development on Information Retrieval, August 2006*.
- M. Koppel, N. Akiva and I. Dagan (2006), [Feature Instability as a Criterion for Selecting Potential Style Markers](#), *JASIST: Journal of the American Society for Information Science and Technology, Volume 57, Issue 11, Pages:1519-1525. July 2007*.
- M. Koppel, J. Schler and E. Bonchek-Dokow (2007), [Measuring Differentiability: Unmasking Pseudonymous Authors](#), *JMLR* 8, July 2007, pp. 1261-1276.
- S. Argamon, M. Koppel, J. W. Pennebaker, and J. Schler (2007), [Mining the Blogosphere: Age, gender and the varieties of self-expression](#), *First Monday*, vol 12(9), September 2007.

- Koppel, M., Schler, J. and Messeri, E. (2008), [Authorship Attribution in Law Enforcement Scenarios](#), "Security Informatics and Terrorism - Patrolling the Web", P. Cantor and B. Shapira (Eds), IOS Press NATO Series.
- Strous, R., Koppel, M., Fine, J., Nahaliel, S., Shaked, G. and Zivotofsky, A. (2008), [Automated Characterization and Identification of Schizophrenia in Writing](#), J. of Nervous and Mental Disorders, 197(8): 585-588
- Koppel, M., Schler, J. and Argamon, S. (2009), [Computational Methods in Authorship Attribution](#), JASIST, 60 (1): 9-26
- Koppel, M. and Diskin, A. (2009), [Measuring Disproportionality, Volatility and Malapportionment: Axiomatization and Solutions](#), Social Choice and Welfare , 33 (2): 281-286
- S. Argamon, M. Koppel, J. Pennebaker and J. Schler (2009), [Automatically Profiling the Author of an Anonymous Text](#), Communications of the ACM , 52 (2): 119-123 (virtual extension).
- M. Koppel, N. Akiva, E. Alshech and K. Bar (2009). [Automatically Classifying Documents by Ideological and Organizational Affiliation](#), Proc. of IEEE Intelligence and Security Informatics, Dallas TX, June 2009.
- Shlesinger, M., M. Koppel, N. Ordan and B. Malkiel (2009). [Markers of translator gender: do they really matter?](#), Copenhagen Studies in Language (38), pp. 183-198
- S. Argamon and M. Koppel (2010), The Rest of the Story: Finding Meaning in Stylistic Variation, in The Structure of Style: Algorithmic Approaches to Understanding Manner and Meaning (S. Argamon, K. Burns & S. Dubnov (eds.)), Springer-Verlag: Berlin, pp. 79-112.
- Koppel, M. Schler, J. and S. Argamon (2011), [Authorship Attribution in the Wild](#), *Language Resources and Evaluation* 45(1) (special issue on Plagiarism and Authorship Analysis)
- M. Koppel, N. Akiva, I. Dershowitz and N. Dershowitz, (2011). [Unsupervised Decomposition of a Document Into Authorial Components](#), *Proceedings of ACL*, Portland OR, June 2011, pp. 1356-1364.
- M. Koppel, J. Schler and S. Argamon (2012), The Fundamental Problem of Authorship Attribution, *English Studies* (Special issue on stylometry and authorship attribution), to appear
- N. Akiva and M. Koppel (2012). [Identifying Distinct Components of a Multi-Author Document](#), *Proceedings of European Intelligence and Security Informatics Conference (EISIC) 2012*, to appear