Project Title: DNA of the Written Word: Using Computer Analysis to Identify Borrowed and Interpolated Material in Medieval History Texts

RFP Category: 1-A  Total Grant Amount Requested from FPDC: $7950

Discipline: Medieval Studies  Sub-Discipline: Computational Linguistics

Project Director:
Sarah Downey, Assistant Professor, Department of English, California University, 724-938-4719, downey@calu.edu

Faculty Status (see definitions below):

☐ Tenured  ☒ Probationary  ☐ Non-Tenure Track

Other Participants:
Undergraduate students at California University:
  Doug Raffle (Computer Science)
  Justin Boyle (Computer Science)
  Veronica Kerekes (History)
  Rich Williams (History)

IRB/IACUC Status: ☐ Approved (IRB # )  ☐ Pending  ☒ N/A

ABSTRACT:
Medieval history texts are often a pastiche of unattributed material from outside sources; in a world without copyright laws, authors freely borrowed content from existing works, and later copyists silently added material to suit their own purposes. Traditional textual analysis has long been used in attempts to isolate and identify such material, but computer-assisted linguistic analysis, which has yielded remarkable results elsewhere, has not yet been applied to medieval historical works. Preliminary tests indicate that lexomics, a system of statistical computer programs modeled on software used for DNA analysis, can identify segments of text whose wording is different from that of the larger works in which they appear. This grant will fund a student-faculty collaboration that will use lexomics software to explore medieval writing in unprecedented ways, showing that new knowledge can still be produced about very old texts, and that computers can change our understanding of history.
The Wheaton Lexomics Group, an interdisciplinary team of professors at Wheaton College in Massachusetts, has pioneered the application of lexomics to medieval texts with the help of an NEH Digital Humanities Startup Grant (http://wheatoncollege.edu/lexomics/). I have been collaborating with the Wheaton Group since May 2009, and we have already had one joint article accepted for publication in *Modern Philology*. Thus far the Wheaton Group has worked almost exclusively on literary texts in Old English; now that I have become familiar with their methods and resources, I intend to use lexomics to analyze major historical works written in medieval Latin.

The Wheaton Group has already demonstrated that lexomic graphing is especially good at identifying sections of a text which come from a different source than the rest of the material. For example, our *Modern Philology* article uses lexomics to show that the wording of a section of the Old English narrative poem *Guthlac A* is statistically different from that of the rest of the poem. We were able to argue that, because this section is almost certainly drawn from a known Latin source while the rest of the poem is more likely from oral tradition or the author’s imagination, the poem either post-dates the Latin source or was substantially altered after the production of the Latin source. This contributes to ongoing scholarly discussion about when *Guthlac A* might have been written and what it can tell us about the development of the English language. Similarly successful applications of lexomics to Old English are detailed in the Lexomic Group’s methodology article for *The Journal of English and Germanic Philology* (listed below under “References”).

In preparation for applying lexomics to medieval Latin historical texts, we have already run preliminary tests on one of the best-known works of medieval history, Bede’s *Ecclesiastical History of the English People*. Because the sources of Bede’s *History* are fairly well-known, we were able to use it as a control. Our initial graphs show exciting promise: the lexomics software has been able to identify and isolate portions of papal letters which Bede famously includes in his *History*, as well as material from another historian (Gildas) and material taken from independent accounts of saints’ lives. This control test indicates that lexomics could provide important evidence for historical texts in which the use of outside sources is debated. If we can identify known excerpts of papal letters in Bede’s *History*, we can likely identify previously unrecognized excerpts in other historical works. Such work, never before attempted, would be of enormous use to the fields of medieval history and Latin language study. Knowing that a text has been altered after its initial composition, or that it is borrowing outside material, can change the way we understand history.
Goals and Objectives

The overarching goal of this project is to produce new knowledge about medieval Latin historical texts, the cultures that produced them, and how we are able to understand those cultures today. In addition, our work will generate interest in, and awareness of, the effectiveness of computational analysis in interpreting premodern documents. Specific objectives associated with these goals include:

- Complete a database of contextual information on all medieval Latin texts which are available online in accessible e-texts (approximately 100 works).
- Generate lexomic graphs of all major medieval Latin historical works which are available online in accessible e-texts (approximately 25-30 works).
- Conduct in-depth research on the most promising 10-15 graphs to explore and explain their patterning, particularly any anomalies that might indicate outside sources.
- Prepare collaborative student-faculty presentations and publications to disseminate our findings. Possible venues include Cal-U’s Academic Excellence Days; the Arizona Undergraduate Conference on Medieval and Renaissance Studies; the 2012 International Medieval Congress in Kalamazoo, MI; the *Keystone Journal of Undergraduate Research*, and appropriate academic journals in the field of medieval studies (e.g. *Modern Philology*).
- Create a framework and momentum for future faculty-student collaboration in lexomics at Cal-U.

Description of Project

I am fortunate to have an enthusiastic group of undergraduate students who are already preparing to collaborate with me on this project. Three History majors (Veronica Kerekes, Josh Solomon, and Rich Williams) and one Computer Science major (Doug Raffle) have been working with me since fall 2010 and are currently participating in a small, unpaid internship which will enable them to acquire the skills they need to be full collaborators in 2011-2012. These four students recently traveled with me to Wheaton College to receive training in lexomics software directly from the Wheaton Lexomics Group; as a result, Veronica, Josh, and Rich are now able to find and prepare electronic texts for analysis, and Doug is able to run the more complicated graph-generating programs in the statistical language R. Doug has also begun tweaking some of the programs’ code for our particular purposes. Over the course of the spring 2011 semester, all of the students will continue to improve their Latin skills and to learn more about how to research medieval Latin historical texts. One of the History majors (Josh Solomon) will graduate this year, and we plan to add another Computer Science major who has expressed interest (Justin Boyle), so in 2011-2012 our research team will consist of two Computer Science majors, two History majors, and me. By summer 2011, the students’ training and preparation will be complete, and we can begin our research in earnest.

Computer analysis requires electronic versions of texts; fortunately for us, many medieval historical works have been digitized in recent years and are now available online at no cost. We will begin our work by assessing the material that is available from websites such as the *Bibliotheca Augustana* and the *Latin Library*, gathering basic background information (author, date, medium, availability of English translations and scholarly editions) for each e-text, and compiling this information into a database which we can use as a resource throughout the project and in future work.
Because this database will not be limited to historical works, but will include all medieval Latin texts available at no cost online, it will be of use to future iterations of the project which may extend to other types of texts (e.g. poetry). More immediately, the database will give us a clear picture of exactly what historical texts are available and suitable for lexomic analysis.

I estimate that about 20-25 major medieval history works, similar to Bede’s *Ecclesiastical History*, are currently available online. Once we have gathered basic contextual information on these texts for the database, we will use our lexomics software to cut these texts into sections and generate statistical graphs. We will scrutinize these graphs closely and collaboratively, using scholarly editions to determine which sections of our texts are known to have outside sources and which are suspected. We will then select those texts whose graphs show promising leads, and begin in-depth investigation of existing scholarly analyses and debates on these texts, looking for ways in which our evidence might contribute new knowledge to the conversation. The students will be full collaborators and will conduct their own original research with my guidance.

We plan to disseminate our findings in spring 2012. I will work together with the students to prepare presentations and/or joint publications. Because all of the students intend to apply to graduate school, I intend for each of them to have at least one presentation and/or publication by the time the project is finished. Our work may result in one large collaborative piece under all of our names, or in shorter pieces which I will co-publish with individual students or pairs of students. In addition, toward the end of the project, I will collaborate with the Wheaton Group on an overarching Latin lexomics methodology article to be submitted to *Speculum*, one of the top journals in the field of Medieval Studies.

**Timeline (Project is planned for 16 months, from May 2, 2011 to August 10, 2012)**

- **2010-2011 school year**: preliminary instruction in Latin language and historical research for students (one hour each week beginning September 2010, currently in progress)
- January 12-15, 2011: travel to Wheaton College for training in lexomics software (*completed*)
- May 2, 2011: **project officially begins**; purchase supplies and equipment
- May 14, 2011: presentation on successful preliminary graphs of Bede’s *Ecclesiastical History* at the International Medieval Congress in Kalamazoo, MI (paper has been accepted)
- May-August 2011: establish framework for database and begin collecting information for entries; modify code of lexomic computer programs for use with Latin texts
- by September 30, 2011: complete at least 50% of database, including entries for 20-25 major historical works
- by October 28, 2011: generate and review lexomic graphs for 20-25 major historical works
- November 2011 – January 2012: intensive research into existing scholarship on texts with most promising graphs
- by February 3, 2012: target specific forms of presentation and/or publication (See “Goals and Objectives” for possible venues)
- February – April 2012: prepare presentations and/or publications; complete at least 75% of database
- May 2012: presentation on progress at the International Medieval Congress in Kalamazoo, MI, probably as part of a joint session with the Wheaton Lexomics Group
- May – August 2012: complete and submit collaborative publications with students; prepare an overarching methodology article with the Wheaton Lexomics Group, to be submitted to *Speculum.*
• by August 10, 2012: complete database of all medieval Latin texts available online; determine avenues for future student-faculty research in lexomics at Cal-U (including maintenance and updating of the database as more texts become available); project ends

**Expected Outcomes**

This project will enable undergraduate students to contribute substantially to their academic fields of study. They will be full collaborators in rigorous and unprecedented research. The students will have opportunities for networking and publication which will benefit their careers enormously, and each student will learn to work as part of a research team. This project will provide momentum for continued involvement of undergraduates in my research at Cal-U.

Our work will result in:

• new knowledge about medieval Latin historical texts and the events they describe
• new, unprecedented applications of computational linguistic analysis
• a database of online medieval Latin texts, which can be used and expanded in future study
• collaborative student-faculty presentations and publications
• a collaborative methodology article with the Wheaton Lexomics Group

Beyond the obvious benefits of print publication, I stand to benefit professionally from this project in other ways. First, by working with the Wheaton Group and presenting yearly papers at the International Medieval Congress, I will be able to maintain contact, on a national level, with scholars both in and out of my field. In addition, this project will help me maintain my knowledge of the Latin language, which is not only a crucial skill for any research in medieval literature or history, but also of continuing importance to me as I teach courses in grammar, linguistics and the history of the English language (ENG 345, ENG 346, and ENG 347). Overall, this project will substantially benefit me, my students, my professional collaborators, and the multiple scholarly disciplines to which it pertains.
## Budget

<table>
<thead>
<tr>
<th>Project Budget</th>
<th>Proposed Grant</th>
<th>University Contribution</th>
<th>Other Revenue Sources</th>
<th>Totals</th>
</tr>
</thead>
<tbody>
<tr>
<td>Salaries/Stipends</td>
<td>$2,000</td>
<td></td>
<td></td>
<td>$2,000</td>
</tr>
<tr>
<td>Student Wages</td>
<td>$4,800</td>
<td></td>
<td></td>
<td>$4,800</td>
</tr>
<tr>
<td>Benefits</td>
<td>XXXXXX</td>
<td>$727</td>
<td></td>
<td>$727</td>
</tr>
<tr>
<td>Honoraria (for consultants)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Supplies</td>
<td>$250</td>
<td></td>
<td></td>
<td>$250</td>
</tr>
<tr>
<td>Equipment</td>
<td>$900</td>
<td></td>
<td></td>
<td>$900</td>
</tr>
<tr>
<td>Operating Expenses</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Travel</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Other (specify)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>TOTALS</strong></td>
<td><strong>$7,950</strong></td>
<td><strong>$727</strong></td>
<td></td>
<td><strong>$8,677</strong></td>
</tr>
</tbody>
</table>

### Budget notes:

Faculty stipend is for summer work in 2011 and 2012.

Student wages are calculated at 120 hours per student at $10/hour, above the minimum wage because these students have already completed extensive, necessary training in lexomics software and medieval Latin in order to collaborate on the project. It is anticipated that 100 student work hours would occur during summer 2011 and 2012. (120 hours X $10 per hour X 4 students = $4,800).

The University will contribute the fringe benefits based on 18% of faculty stipend (18% X $2,000 = $360) and 7.65% of student wages (7.65% X $4,800 = $367).

Supplies in the amount of $250 include printer paper, toner, and similar materials to support project activity.

Equipment in the amount of $900 consists of: 1) a dedicated laptop on which lexomics software can be installed, and which the students can check out during assigned work hours (necessary because lexomics software cannot be installed on other University computers), and 2) an external hard drive for backing up project data.
References


Wheaton Lexomics Group website:  http://wheatoncollege.edu/lexomics/
Sarah Downey  
Department of English email: downey@calu.edu  
California University of Pennsylvania office: 724-938-4719  
250 University Avenue fax: 724-938-1515  
California, PA 15419 home: 412-361-1673

Education:  
PhD, Medieval Studies: Centre for Medieval Studies, University of Toronto 1999 – 2004  
BA, Latin (summa cum laude): University of the South 1995 – 1999

Employment:  
Assistant Professor (tenure-track) 2009 – present  
Department of English, California University of Pennsylvania  
Lecturer (full-time) 2005 – 2009  
Department of English and Department of Classics, University of Tennessee Knoxville  
Adjunct Assistant Professor 2004  
School of Architecture, University of Waterloo  
Instructor 2004  
Department of English, University of Toronto

Publications:  
in print:  
multiple reviews for Notes & Queries and Choice, encyclopedia entries on St. Guthlac

accepted for publication:  

in preparation:  
with Paul Crawford (Cal-U Department of History and Political Science): The Passion of Renaud. translation of a Latin biography of Renaud de Châtillon, a Crusades leader; to be published as part of a collection of sources about Renaud, edited by Bernard Hamilton.  
with Maura Lafferty (University of Tennessee Department of Classics) and the Wheaton Lexomics Group. “Lexomic Studies of Medieval Latin Texts.” to be submitted to Speculum.  
Selected research presentations:

“Computational Linguistics and Old English Demons”  
Faculty Professional Development Research Subcommittee Luncheon  
California University of PA  
April 13, 2010

“Beowulf and Tolkien in Freshman Comp”  
Medieval and Renaissance Teaching Conference  
Carson Newman College, Jefferson City, TN  
October 2009

“A Trio of Guthlacs: The South English Legendary Lives and the Guthlac Tradition”  
41st International Congress on Medieval Studies  
Western Michigan University, Kalamazoo, MI  
May 2006

Canadian Society of Medievalists Annual Congress  
University of Western Ontario, London, ON  
May 2005

“Excessive Fasting and the Old English Prose Life of St. Guthlac”  
40th International Congress on Medieval Studies  
Western Michigan University, Kalamazoo, MI  
May 2005

“Qui frequentatur ab hominibus frequentari ab angelis nequit: The Tradition of St. Guthlac’s Solitude” (winner of Colloquium prize)  
Sewanee Medieval Colloquium  
University of the South, Sewanee, TN  
April 2005

“Formulas in Felix’s Vita Sancti Guthlaci”  
39th International Congress on Medieval Studies  
Western Michigan University, Kalamazoo, MI  
May 2004

Selected university service:

Division Editor (English), Keystone Journal of Undergraduate Research (statewide)  
2009 – 2011

Member, FPDC Research Subcommittee  
2010 – 2011

Member, planning committee for 2012 PASSHE Summer Honors Program in the UK  
2010 – 2011

Co-director, “Warfare, Wounds and Disease in the Middle Ages”  
(October 20-22 public event, including performance of Beowulf by Benjamin Bagby)  
2010

Accompanied seven Cal students to the International Congress on Medieval Studies in Kalamazoo, MI  
May 13–16, 2010

Accompanied three English majors presenting papers at the English Association of Pennsylvania State Universities undergraduate conference  
April 8–9, 2010

Advisor for Justin Grimm, “Attila the Hun’s Role in the Fall of the Western Roman Empire through the Displacement of Germanic Tribes” (honors thesis)  
2009 – 2010

Advisor for Justin Piper, “Interpretations of The Lord of the Rings” (honors thesis)  
2009 – 2010