White Paper Report

Report ID: 102837

Application Number: PR5011211

Project Director: Michael Drout (mdrout@wheatoncollege.edu)

Institution: Wheaton College

Reporting Period: 4/1/2011-8/31/2013

Report Due: 11/30/2013

Date Submitted: 11/26/2013

- White Paper
- PR-50112-11
- Lexomic Tools and Methods for Textual Analysis: Providing Deep Access to Digitized Texts
- Michael D.C. Drout, Michael J. Kahn, Mark D. LeBlanc (co-PIs)
- Wheaton College, Norton, MA
- November 20, 2013

This grant allowed the Wheaton Lexomics Research Group to do extensive and long-term research over three summers and two winters (beyond the original scope of the grant). During the summers of 2011 and 2012, 3 full professors from Wheaton (the co-PIs), an associate professor from Wheaton, 3 visiting professors from other institutions, 15 undergraduate students from Wheaton and 3 undergraduates from other institutions worked full-time both to develop new software for textual analysis and then developed methods for using the software on a variety of texts ranging from the Anglo-Saxon period to the 20th-century Harlem Renaissance. The research continued in the summer of 2013 with the participation of 2 full professors from Wheaton, 8 additional undergraduates and 1 visiting professor. Supplemented by other funds from Wheaton College and from faculty members' personal research budgets, the grant supported presentation of our research at multiple conferences on both sides of the Atlantic and in fields as diverse as Harlem Renaissance studies, digital humanities and Anglo-Saxon. Detailed explanatory materials, software downloads and on-line tools are all available at http://lexomics.wheatoncollege.edu

Developed at Wheaton College with the partial support of a previous grant from the National Endowment for the Humanities, "lexomic" methods are an outgrowth of work on computational stylometry by John Burrows, David Hoover and others. Enabled by the recent proliferation of high-quality digital editions, lexomic analyses employ computer-assisted statistical techniques to identify patterns, which are then interpreted using traditional literary methods. Over the course of the grant we developed first a suite of software tools for textual analysis and then integrated these into the single Lexos Integrated Workflow that is available for use on the project website at http://lexos.wheatoncollege.edu. We also created detailed explanatory and instructional materials (videos and manuals) available on our website. As part of an "iterate and test" approach, the software evolved in tandem with both the techniques that employed it and the particular research questions we were investigating.

After further evaluation and suggestions from Prof. Scott Kleinman, California State University, Northridge, we began to more fully integrate Lexomics with the Digital Humanities scholarly community as well as reaching out to scholars and, in particular, graduate students in traditional fields who might benefit from the methods and tools. This outreach culminated in a hands-on graduate workshop at the International Society of Anglo-Saxonists conference in Dublin in July 2013, in which we taught 25 graduate students how to use lexomics tools. Our presentation was the high point of the digital humanities “pre-conference” workshop attended by these students, who are now using lexomic methods as part of their research.

The success of the workshop was enabled, in great part, by our having gone beyond the original work plan in the grant (which was completed by the end of summer 2012) to create the Lexos Integrated Workflow, which bundled all of the previously created tools together, enabling scholars to use them through a simple and convenient web interface (rather than the more complex set of programs that were called for in the original grant). Lexos made the tools so easy to use that graduate students at the workshop were able to pick them up in less than an hour, and we in fact made a significant discovery in front of the live audience (co-PI Drout had uploaded the wrong text into a demonstration and was pleasantly surprised to see the unexpected correlation between the distribution of an arbitrarily selected conjunction and the textual history of an Anglo-Saxon poem).

Discoveries enabled by the lexomics software and the methods developed alongside it have led to major publications and also to work that is currently under consideration or in progress. Visiting scholars have used the tools and techniques to investigate texts ranging from Zora Neale Hurston’s and Langston Hughes’ play *Mule Bone* to Shakespeare’s *The Two Noble Kinsmen*, the Anglo-Saxon poem *Beowulf*, *Reykdæla Saga* and *Víga-Glums Saga* in Old Norse, and Latin texts by the Geoffrey of Monmouth, Alan of Lille and the Venerable Bede.

Software Development: Three major tools were developed for textual analysis: Scrubber, which processes texts to make them analyzable by computer; DiviText, which divides texts into segments; and TreeView, which produces branching diagrams (dendrograms) of vocabulary distribution in the texts. These programs, which are all freely distributed both directly on the project website and through

GitHub, were then integrated into the web-interfaced Lexos Integrated Workflow, which allows users to employ the tools without having to download or install them or run them from the command line. All the software is extensively documented, and the website includes detailed instructional videos and instruction manuals that explain not only how to use the tools, but how they work and why researchers might want to employ them in textual analysis.

We have published in major journals, communicated through the Old English newsletter, and presented our research at the Northeast Modern Language Association conference, the International Medieval Congresses of 2011 and 2012, the International Society of Anglo-Saxonists conference of 2013, Digital Humanities 2013, and at guest lectures and presentations at a multitude of institutions.

Because we changed the website after the development of the Lexos Integrated Workflow, we have two sets of data on audience. For the first two years of the website, 2011 and 2012, we had 15,000 unique visitors who stayed on the site an average of 1:34 minutes. But with the advent of Lexos, we had 544 visitors in only 3 months, viewing over 6000 page views and staying on the site a remarkable average of nearly 9 minutes. This long duration shows that people are using the tools rather than just perusing the site. Visitors come from a wide range of countries, mostly in North American and Western Europe, but also from Iran, Libya, the UAE and Australia.

The lexomics project was evaluated twice: once by our advisory board at the International Medieval Congress at Western Michigan University in 2012, and once by our consultant, Prof. Scott Kleinman, in July 2013. Our evaluation by the advisory board focused on the intellectual validity of the approach, ease-of-use of the tools, interaction with the Dictionary of Old English website and dissemination to the scholarly community. The board was primarily concerned that we might have moved down a path towards convincing ourselves of the utility of our techniques without explaining to a wider audience why we had confidence in them. After the board meeting we changed our documentation and explanatory focus from the “In the Margins” approach (in which a variety of wiki-like materials would be linked to the tools on our website) to a more comprehensive set of instructional videos and texts. This effort was at times extremely time-consuming, using much of our energy in the summer of 2012, but it turned out that the board’s advice was wise: not only did we better communicate the intellectual underpinnings of the techniques to a wider audience, but having to explain the material in a coherent fashion improved both our techniques and our understanding of them.

Our evaluation by Scott Kleinman was more technical and focused on both the “back end” of the tools (how they worked in software terms, what their capabilities were) and in finding ways to better engage the project with the conversations of the wider Digital Humanities community. In response to his evaluation, we added external tools to the Lexos Integrated Workflow “Analysis” page, incorporated insights from other Digital Humanities projects into both our software and our techniques, and revised the user interface for the tools.

We have received an enthusiastic response every time we have been able to demonstrate the Lexomics tools to audiences. Scholars immediately want to try the tools, especially since they are now embodied in the Lexos Integrated Workflow. We have hosted workshops at Wheaton, given lectures at other institutions, and ran that above-mentioned workshop for 25 graduate students at the International Society of Anglo-Saxonists bi-annual meeting in Dublin in July 2013.

In addition to the fortuitous live discovery at the Dublin workshop described on page 2, there were a number of other moments when we knew we were on to something good:

- At the Dublin workshop, a graduate student from Glasgow had questions about the Anglo-Saxon poem “The Descent into Hell.” Using the Lexos tools, we were able to “scrub” an electronic version of her text, cut it into segments and produce a dendrogram in only 5 minutes. This process, even with our previous iteration of tools, would have taken an hour, and it would have been the work of perhaps a week with the very first approaches we developed. When we realized that we could do all of this work in almost no time at all, and that we took for granted that we could scrub the text in multiple ways, cut it into segments of various sizes and immediately analyze it not only in a dendrogram, but with an entirely new technique, a “rolling window analysis,” we knew we had produced a significant and useful set of tools and methods.
- Our previous work in “Scrubbing” texts had been complicated by the use of two interchangeable symbols, þ and ð, in Anglo-Saxon orthography. We noticed that only two particular segments in a dendrogram seemed to be affected by “consolidating” all ð to þ (so as to count words more accurately). From this small observation arose the technique we are now calling “Theta-Analysis” or “Rolling Window Analysis,” which allows us to produce graphs of the changing frequencies of the two symbols throughout at text. The most significant inflection point in one of these graphs turned out to be related to a new archeological find that was announced at the Dublin ISAS conference: a

recently discovered object contains a short runic inscription that matches three lines of the poem *Daniel*, three lines precisely coincident to the change in þ/ð ratio in the text.

- The Old Norse texts *Víga-Glums Saga* and *Reykdæla Saga* share a chapter about a character named Víga-Skúta. Scholars have long debated the priority of the two texts. An elegant experiment was able to show definitively that the vocabulary distribution of shared chapter is much closer to that of *Víga-Glums Saga*, settling a long-standing question (and the key experiment was performed by one of our undergraduate research partners!).
- The Venerable Bede's *Ecclesiastical History* contains a letter purported from the Abbot Ceolfrith. Scholars have long argued that, based on some similarities to other parts of the *Ecclesiastical History*, the letter is actually by Bede. Our research showed that critics have it backwards, and that one part of the *History* that Bede does not attribute is nevertheless likely to have been written by Ceolfrith. Further traditional research strongly supports the idea that using Ceolfrith by name in this section of the text would have been politically difficult for Bede, explaining why the evidence is consistent with him borrowing the text but not attributing it.

Continuation of the Project

Lexomics has now become so popular and has generated so much interest that it would continue even if we did not want to continue working on it. Researchers are using the tools and applying them in ways we had not previously imagined (For example, it turns out that the Scrubber part of the Lexos Integrated Workflow has wide application to many other Digital Humanities projects that do not use the other lexomic techniques). But we are in fact going to continue the development of lexomic methods, in our classes at Wheaton, as part of summer research with students, and, hopefully, as part of a new NEH Digital Humanities Start-up grant for which we have applied.

Our goal now is to expand the applicability of the methods from English to other languages and from the Roman alphabet to other writing systems. In the near future, we plan is to add localization for languages such as Spanish and Japanese, and to modify the lexomic methods to explore their use for the study of Spanish and Japanese literature.

Lexomic methods are now a part of at least three undergraduate classes at Wheaton College and are also being used at California State University, Northridge and California University of Pennsylvania. We have had a steady stream of students who have heard about the techniques and want to try them (WARNING: creating and interpreting dendrograms is *very* addicting!)

The success of the project—undergraduate students made significant *discoveries* (not normally an accomplishment associated with literary studies)—has led to our institution supporting further research both through funding and the provision of laboratory space and technical support. Lexomics is now a high-profile success story at Wheaton.

Recent Talks and Workshops

- Drout, M.D.C., LeBlanc, M.D., Neal, R. '14, Berger, R. '15, Hitotsubashi, N. '13, Smith, L. '14. (2013). **Graduate Workshop: Using Lexomics Tools.** Presented at the *International Society of Anglo-Saxonists (ISAS 2013)*, Dublin, Ireland, July 28, 2013.
- LeBlanc, M.D., Drout, M.D.C., Kahn, M., Herbert, A. '14, Neal, R. '14 (2013). **"Lexomics: Integrating the research and teaching spaces."** Presented at and published in proceedings of *Digital Humanities 2013*, University of Nebraska–Lincoln, July 2013: pages 274-276.
- Drout, M.D.C. and LeBlanc, M.D. (2013). "Mule Bone 2.0," in Session 14.05, the Literary Interventions of the Digital Humanities, **44th Annual NeMLA Convention**, March 21-24, 2013, Tufts University, Boston.
- Christian, S. (collaborator) and La Brie, C. (2013). "Mule Bone 2.0," in Session 5.08, Make it New: Approach for Teaching the Harlem Renaissance, **44th Annual NeMLA Convention**, March 21-24, 2013, Tufts University, Boston.
- M. Drout. "Lexomics: New Digital Methods for Old English Texts." Quod Libet. Cornell University. March 15, 2012.
- [upcoming] M. Drout. "The Persistence of Influence and How to Detect It." Featured Lecture, *The Presents of the Past Symposium*, Texas A&M University, April 4–5, 2014.

Publications

- Boyd, P., Drout, M.D.C., Hitotsubashi, N., Kahn, M., LeBlanc, M.D. and Smith, L. (in press, 2014). "Lexomic Analysis of Anglo-Saxon Prose: Establishing Controls with the Old

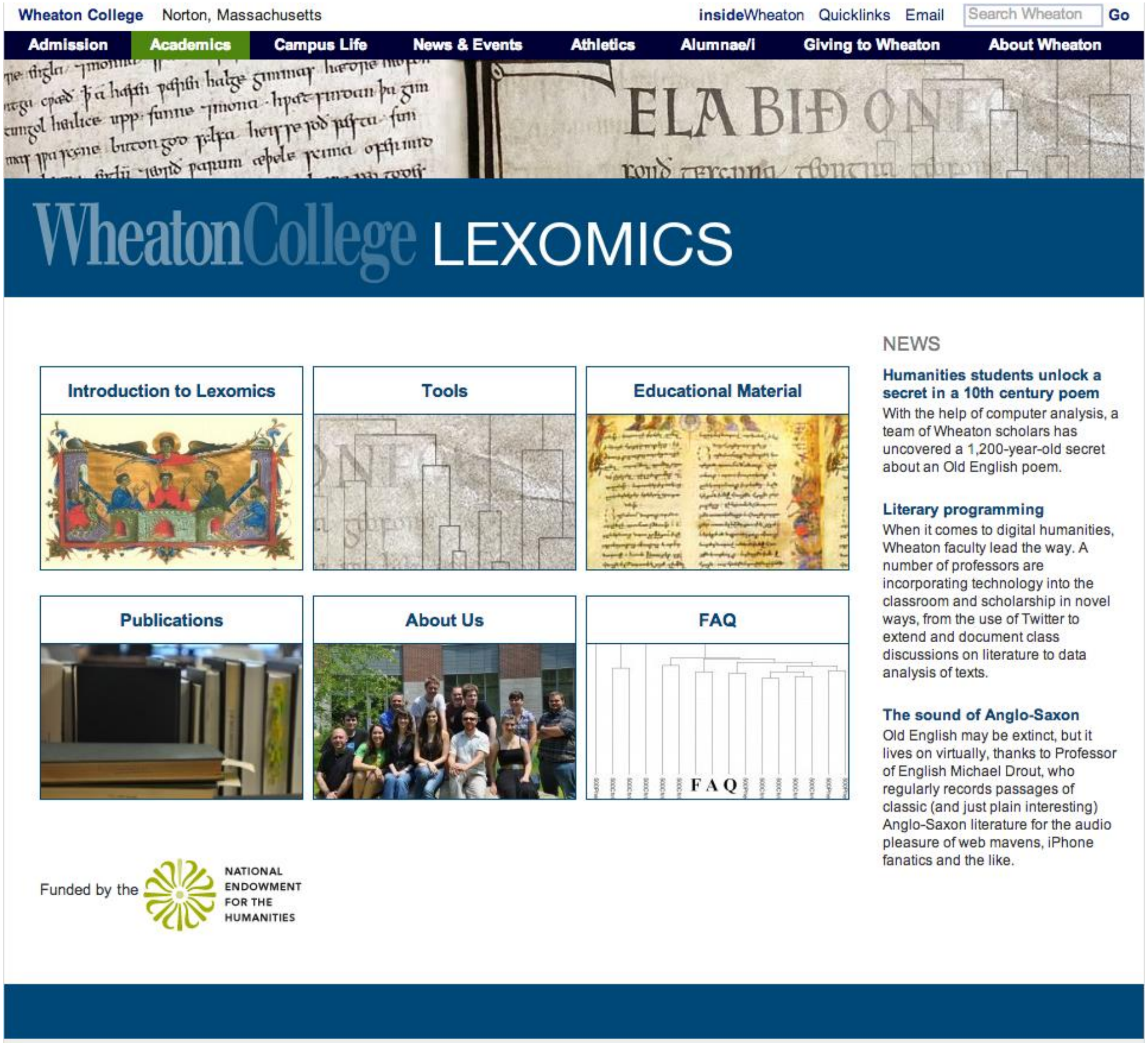
English Penitential and the Old English translation of Orosius." *Revista de la Sociedad Española de Lengua y Literatura Inglesa Medieval (SELIM)*.

- Drout, M.D.C. (2013). ***Tradition and Influence in Anglo-Saxon Literature: An Evolutionary, Cognitivist Approach*** (New York: Palgrave, 2013), pages 47-82.
- [under review] Sarah Downey (collaborator), Michael D.C. Drout, Veronica Kerekes and Douglas Raffle, "Lexomic Analysis of Medieval Latin Texts", *Journal of Medieval Latin*.
- Downey, S. (collaborator), Drout, M.D.C., Kahn, M., LeBlanc, M.D., **"'Books Tell Us': Lexomic and Traditional Evidence for the Sources of *Guthlac A*."** *Modern Philology* 110 (2012):153-181.
- Drout, M.D.C., Kahn, M., LeBlanc, M.D., Nelson, C. '11 "**Of Dendrogrammatology: Lexomic Methods for Analyzing the Relationships Among Old English Poems**," *Journal of English and Germanic Philology*, v110(3), July 2011, 301-336.
- Drout, M.D.C., Kahn, M., LeBlanc, M.D., Jones, A. '11, Kathok, N. '10, and Nelson, C. '11. **"Lexomics for Anglo-Saxon Literature."** *Old English Newsletter*, 2010.
- LeBlanc, M.D., Gousie, M. and Armstrong, T. (March 2010).**Connecting Across Campus.** Proceedings of the 41st SIGCSE Technical Symposium on Computer Science Education, Milwaukee, WI.
- LeBlanc, M.D. **Computing for Poets**. Presented at SIGCSE 2010 – The Technical Symposium on Computer Science Education, Milwaukee, WI, March 12, 2010.
- [in press] Phoebe Boyd, Michael D.C. Drout, Namiko Hitotsubashi, Michael J. Kahn, Mark D. LeBlanc and Leah Smith. "Lexomic Analysis of Anglo-Saxon Prose: Establishing Controls with the Old English Penitential and the Old English translation of Orosius." *Revista de la Sociedad Española de Lengua y Literatura Inglesa Medieval (SELIM)* 19 (2014).
- [in press] Michael D.C. Drout, Namiko Hitotsubashi and Rachel Scavera. "The Evolution of J.R.R. Tolkien's Túrin Story." *Tolkien Studies*.
- [under consideration] Sarah Downey, Michael D.C. Drout, Veronica Kerekes and Douglas Raffle. "Lexomic Analysis of Medieval Latin Texts, *Journal of Medieval Latin*.
- [in progress] Michael D.C. Drout, Yvette Kisor, Elie Chauvet, Allison Dennett, Natasha Piirainen and Leah Smith. "Lexomic Analysis of *Beowulf*."
- [in progress] Elie Chauvet, Michael D.C. Drout, Michael J. Kahn, Mark D. LeBlanc, and Leah Smith "Lexomic Analysis of Poems Signed by, Attributed to and Related to Cynewulf."

- [in progress] Rosetta Berger and Michael D.C. Drout. “A Reconsideration of the Relationship Between *Víga-Glúms Saga* and *Reykdæla Saga*: New Evidence from Lexomic Analysis.”
- [in progress] Michael D.C. Drout and Elie Chauvet, “Visual Representation of the Ratio of þ to þ+ð: A New Tool for the Investigation of Old English Textual History.”

2. Appendices

Screenshots of the Lexomics website



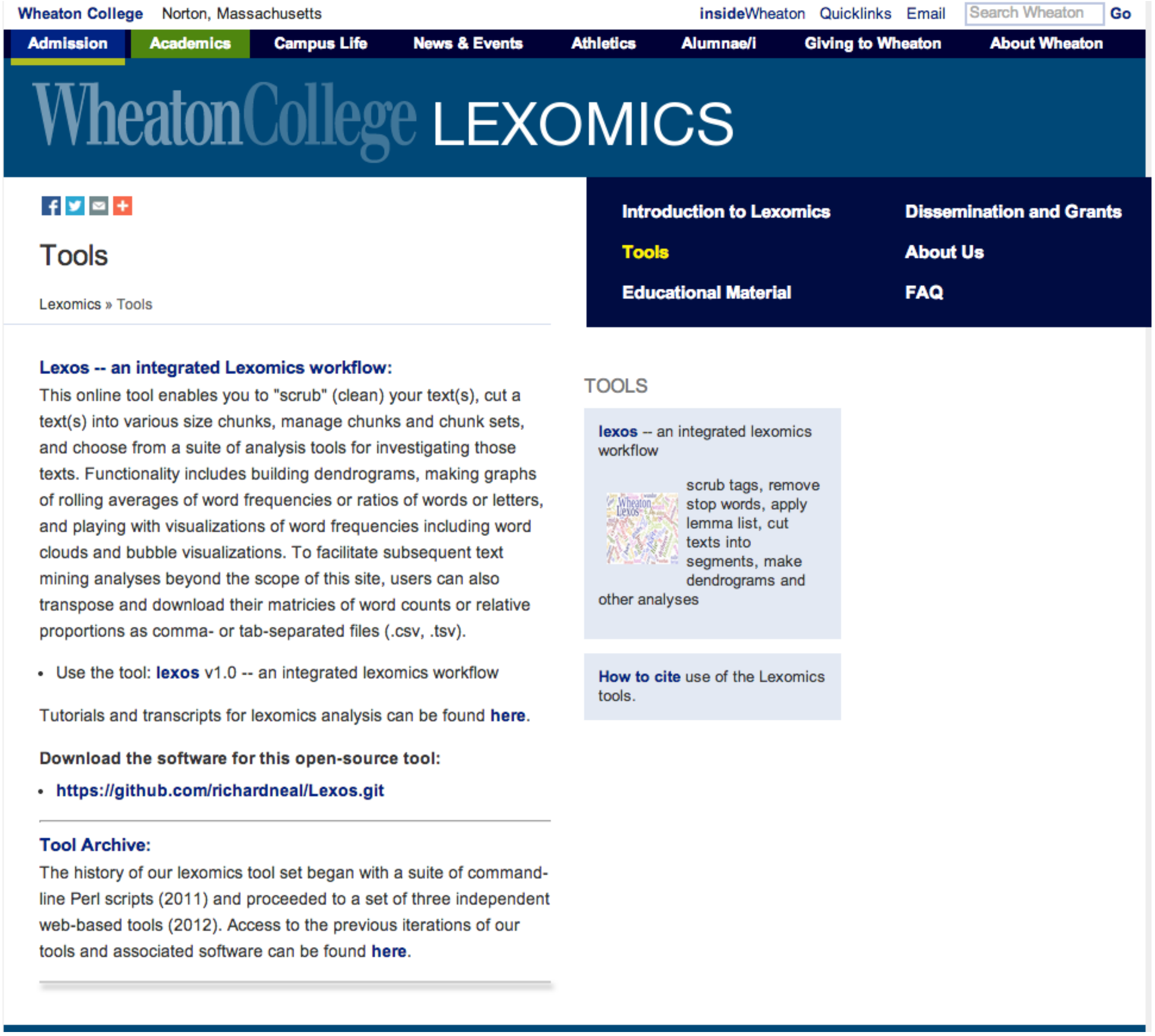
NEWS

Humanities students unlock a secret in a 10th century poem
With the help of computer analysis, a team of Wheaton scholars has uncovered a 1,200-year-old secret about an Old English poem.

Literary programming
When it comes to digital humanities, Wheaton faculty lead the way. A number of professors are incorporating technology into the classroom and scholarship in novel ways, from the use of Twitter to extend and document class discussions on literature to data analysis of texts.

The sound of Anglo-Saxon
Old English may be extinct, but it lives on virtually, thanks to Professor of English Michael Drout, who regularly records passages of classic (and just plain interesting) Anglo-Saxon literature for the audio pleasure of web mavens, iPhone fanatics and the like.

Funded by the NATIONAL ENDOWMENT FOR THE HUMANITIES



Tools

Lexos -- an integrated Lexomics workflow:

This online tool enables you to "scrub" (clean) your text(s), cut a text(s) into various size chunks, manage chunks and chunk sets, and choose from a suite of analysis tools for investigating those texts. Functionality includes building dendrograms, making graphs of rolling averages of word frequencies or ratios of words or letters, and playing with visualizations of word frequencies including word clouds and bubble visualizations. To facilitate subsequent text mining analyses beyond the scope of this site, users can also transpose and download their matricies of word counts or relative proportions as comma- or tab-separated files (.csv, .tsv).

- Use the tool: **lexos** v1.0 -- an integrated lexomics workflow

Tutorials and transcripts for lexomics analysis can be found **here**.

Download the software for this open-source tool:

- **https://github.com/richardneal/Lexos.git**

Tool Archive:

The history of our lexomics tool set began with a suite of command-line Perl scripts (2011) and proceeded to a set of three independent web-based tools (2012). Access to the previous iterations of our tools and associated software can be found **here**.

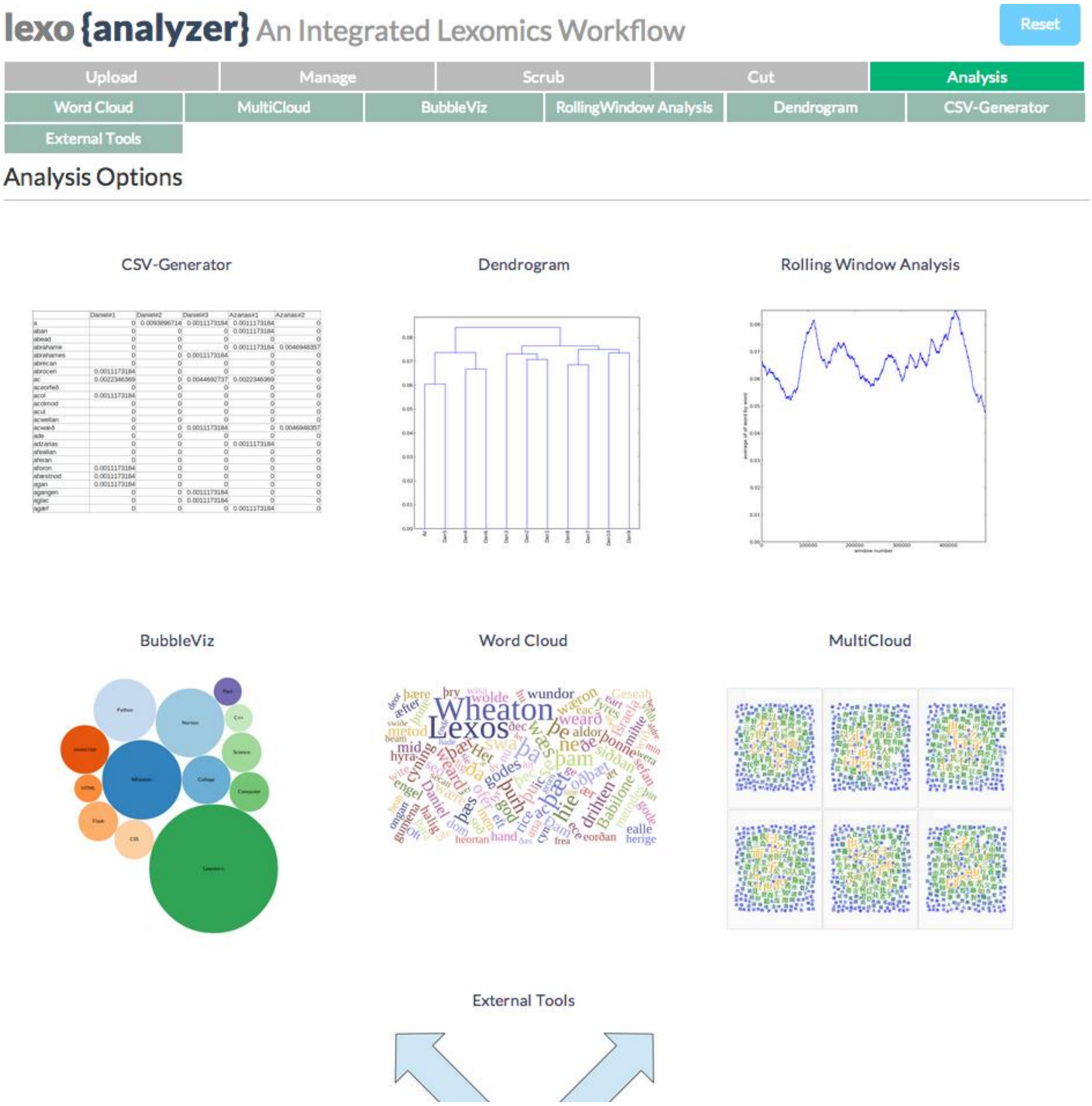
TOOLS

lexos -- an integrated lexomics workflow

scrub tags, remove stop words, apply lemma list, cut texts into segments, make dendrograms and other analyses

How to cite use of the Lexomics tools.

representative samples of completed work,



	Daniel#1	Daniel#2	Daniel#3	Azarias#1	Azarias#2
a	0	0.0093896714	0.0011173184	0.0011173184	0
aban	0	0	0	0.0011173184	0
abead	0	0	0	0	0
abrahame	0	0	0	0.0011173184	0.0046948357
abrahames	0	0	0.0011173184	0	0
abrecan	0	0	0	0	0
abrocen	0.0011173184	0	0	0	0
ac	0.0022346369	0	0.0044692737	0.0022346369	0
aceorfeð	0	0	0	0	0
acol	0.0011173184	0	0	0	0
acolmod	0	0	0	0	0
acul	0	0	0	0	0
acwellan	0	0	0	0	0
acwæð	0	0	0.0011173184	0	0.0046948357
ade	0	0	0	0	0
adzarias	0	0	0	0.0011173184	0
afeallan	0	0	0	0	0
aferan	0	0	0	0	0
aforon	0.0011173184	0	0	0	0
afæstnod	0.0011173184	0	0	0	0
agan	0.0011173184	0	0	0	0
agangen	0	0	0.0011173184	0	0
aglac	0	0	0.0011173184	0	0
agæf	0	0	0	0.0011173184	0