Did Gaius Julius Caesar Write De Bello Hispaniensi?

A Computational Study of Latin Classics

Olivia R. Zhang*, Trevor Cohen** & Scott McGill***
*St. John's School, **Rice University, ***University of Texas Health Science Center at Houston

This project addresses a two-millennium old mystery surrounding the authorship of ancient Latin war memoirs attributed to Gaius Julius Caesar, using methods of distributional semantics, which derive estimates of the similarity between units of text from the distributional statistics of the terms they contain. Of these war memoirs, the Civil War *has been confirmed to be Caesar's work, as have the first seven of the eight chapters of the* Gallic War*, with the eighth authored by Caesar's lieutenant Aulus Hirtius. However, the authorship of the* African War*,* Alexandrine War*, and* Spanish War*, though attributed to Caesar, is still under debate. Methods of distributional semantics derive representations of words from their distribution across large amounts of text, such that words that occur in similar contexts will have similar representations. These representations can then be combined to model larger units of text, such as chapters and entire books. The current work uses one such method, Random Indexing, to calculate the similarity between chapters or books. The results show that the* Gallic War*'s eighth chapter is significantly different both from its other seven chapters and from the* Civil War*, verifying the utility of Random Indexing models as a means to detect different Latin authorships. The* African War*,* Alexandrine War*, and* Spanish War *are also significantly different from those chapters acknowledged to be authored by Caesar, suggesting that he did not write these three.*

Zhang, Olivia R., Trevor Cohen & Scott McGill. "Did Gaius Julius Caesar Write *De Bello Hispaniensi*? A Computational Study of Latin Classics Authorship". *HUMAN IT* 14.1 (2018): 28–58. © The authors. Published by University of Borås.

Furthermore, the African War*,* Alexandrine War*, and* Spanish War *are different from each other, and from the* Civil War *and* Gallic War*, suggesting that they were each written by a different author. This project demonstrates the value of methods of distributional semantics in Classics research. Of note, these methods do not require manual selection or engineering of features for similarity measures, which distinguishes them from the majority of prior statistical and machine-learning methods of authorship attribution. The implications of distributional semantics for digital humanities and related problems such as the evolution of languages over time and plagiarism detection are discussed.*

Keywords: authorship attribution, Caesar, Classics, computational linguistics, distributional semantics, Latin

Authorship attribution for ancient writings is of interest to many historians, literary scholars, linguists, and psychologists. The task of author attribution is to find a reliable method for analyzing a text to determine its authorship. In classical studies, authorship attribution has been traditionally based on historical and literary analyses (e.g., Holmes 2009), which are often subjective and not reliable. Computational methods of authorship attribution, which are based on objective measures of linguistic features and properties, utilize statistical data – such as counts of the frequencies of manually selected "function" words (words with primarily grammatical function), sentence length distributions, and frequencies of punctuation marks – to attribute a text to an author (for early studies, see Mendenhall 1887 and Mascol 1888a/b; for a more recent influential study, see the analysis of "The Federalist Papers" by Mosteller and Wallace 1964; and for comprehensive reviews, see Juola 2008 and Koppel *et al.* 2009).

A prerequisite for the application of such statistical methods is the development of a metric of similarity between texts, such that texts written

by the same author are more similar to each other than to those by other authors. In other words, the statistical properties are akin to authorship "fingerprints" that can be used to identify specific authors. For example, Relative Vocabulary Overlap (RVO) was proposed as a measure of the degree to which two texts draw from the same vocabulary (Ule 1982). In a seminal study of authorship attribution that concerned the authorship of the Federalist Papers, Mosteller and Wallace (1964) used synonym pairs (e.g., "big" and "large") to see whether authors preferred to use one or the other. They also analyzed the frequencies with authors who utilized a set of manually selected "function words" (such as conjunctions, prepositions, and articles), words that carry little meaning but that may vary in the frequency with which they are utilized in a manner characteristic of a particular author. Through the analyses of these and other features, they were able to determine three different authors (John Jay, Alexander Hamilton, and James Madison) under one penname "Publius" for a series of newspaper essays published between 1787 and 1788. The advancement of computational methods of text analysis has provided new approaches to the problem of authorship attribution (for a review, see Stamatatos 2009; see also Koppel *et al*. 2009; Juola 2008; Savoy 2013; Seroussi 2014). In addition to authorship, these computational methods have also been used to study other aspects of writing style, in order to characterize the influence on writing of gender (e.g., Koppel *et al*. 2002; Argamon *et al*. 2003), age (Burger & Henderson 2006; Schler *et al*. 2006), and native language of the author (Koppel *et al*. 2005). These methods can be broadly characterized as either *statistical* methods, in which a distance metric between texts is derived from the difference in their allotment of selected features (such as function words), or *supervised machine learning* methods, in which a classifier is trained on samples of text from different authors, often with selection of features similar to those utilized by statistical models.

In the area of authorship attribution, several recent statistical approaches developed the concept of a "stylome" – a specific set of measurable

traits or features that can be used to uniquely identify a given author (van Halteren *et al*. 2005). For example, Principal Component Analysis (PCA) has been used to derive a two-dimensional visualization of a set of authors (rather than texts) from a multidimensional space of manually selected linguistic features, such as the top 50 most common word types (Burrows 2002). Other methods to infer structure from sets of pairwise distance relations, such as Multidimensional Scaling (MDS) and Hierarchical Cluster Analysis, have also been applied in this domain (Juola 1997, 2007, 2008). Statistical methods used to estimate the distance between texts include the Delta method (Burrows 2002) and the Chi-Square method (Grieve 2007), amongst others (see Juola 2008, for an extensive review).

Machine learning methods have also been applied to authorship attribution. Several early methods used various models of neural networks with small sets of function words as features (e.g., Matthews & Merriam 1993; Merriam & Matthews 1994). More recent studies used wider and larger varieties of features (e.g., Graham *et al*. 2005). Machine learning classifiers that have been applied to this problem include Linear Discriminant Analysis (Baayen *et al*. 2002), Naïve Bayes classifers (e.g., Altheneyan & Menai 2014; Peng *et al.* 2004; Mitchell 1997), and a k-nearest neighbor approach (e.g., Zhao & Zobel 2005).

A commonality to most of these statistical and machine-learning based approaches is that they involve the manual selection of features (such as function words or synonym sets) deemed by the modeler to be of value for authorship attribution. Estimates of similarity are derived from the extent to which these hand-selected features are present in each text, often using a geometric approach in which texts are viewed as points in a vector space with dimensions corresponding to the selected features, such that similar texts are relatively close to one another in this space.

Geometrically motivated methods of distributional semantics (for reviews, see Cohen and Widdows 2009 and Turney 2010) estimate the similarity between terms, and larger units of text, in a similar manner.

However, these distributional semantics methods generally do not employ manual feature selection or engineering[1] (other than exclusion of very frequent or very infrequent terms) and are commonly applied to estimate *semantic similarity*, such as similarity in *meaning* between a pair of terms, or between paragraphs.

At first glance, these distributional semantics methods do not appear well suited to a problem that has been addressed by selecting semantically agnostic function words as features. Nonetheless, in our previous work one such method, known as Random Indexing, was successfully applied to a problem of authorship attribution concerning the Synoptic Gospels (Widdows & Cohen 2009). Furthermore, Seroussi (2014) and Savoy (2013) subsequently evaluated the application of another method of distributional semantics, Latent Dirichlet Allocation (LDA, Blei 2012), to attribute authorship – albeit in cases in which both topic and author varied across documents. Should these approaches be generalizable, the application of methods of distributional semantics for this purpose would present a desirable alternative to approaches requiring manual feature engineering, especially when approaching text in languages other than English, where in many cases sets of terms and other features that are of value for authorship attribution have yet to be constructed. Motivated by the potential utility of a simple, language-agnostic approach to authorship attribution, the current study extends our previous work by evaluating Random Indexing based authorship attribution of ancient Latin texts.

In the current study we use distributional semantics to study the authorship of ancient Latin Literature. The Latin language, especially ancient Latin, has not been as extensively studied as modern languages such as English. Distributional semantics, which does not require manual pre-selection and determination of features for analysis of authorship, has advantages in the analysis of rare, ancient, or forgotten languages – as we shall demonstrate, features of interest such as function words emerge automatically as a result of the analysis. Methods of distributional semantics

are relatively recent computational approaches to automated analysis of natural language text and have wide applications including information retrieval (Deerwester *et al.* 1990), automated grading of content based essays (Landauer *et al.* 1997), and the identification and resolution of ambiguous terms (Scheutze 1998). Unlike other computational methods that focus on overlapping words or phrases, methods of distributional semantics aim to estimate the semantic relatedness between documents by representing the "latent" concepts underlying words as they appear in these documents (hence the name of "Latent Semantic Analysis", Deerwester *et al.* 1990, a widely used distributional semantic method). Of particular importance to the current study, Random Indexing (Kanerva, Kristofferson & Holt 2000), a relatively recent and highly scalable method of distributional semantics, has been successfully applied to study the authorship of certain books of the New Testament (Widdows & Cohen 2009). The aims of the current study are: (1) to see whether distributional semantics can provide new insight into the aforementioned authorship debate around the works by Gaius Julius Caesar; (2) to determine if methods of distributional semantics, developed for the English language and primarily applied to it, can be applied to study ancient Latin texts as a language-agnostic method; and (3) to establish the utility of a method that does not require manual pre-selection of features for similarity comparisons, in the context of authorship attribution.

To these ends, five war memoirs associated with Caesar were selected for evaluation. Of these, the *Civil War* has been confirmed to be Caesar's work, as well as the first seven of the eight chapters of the *Gallic War*, the eighth being written by Aulus Hirtius. The authorships of the three other commentaries, the *African War*, *Alexandrine War*, and *Spanish War*, though attributed to Caesar, are still under debate (Carter 1997; Conte 1994; Hall 1996; Storch 1977). For example, in the *Introduction (p. xxxii–xxxvi)* to the English translation of Caesar's *Civil War,* Carter

summarized the debates by classical scholars on Caesar's authorship in terms of the contents, styles, and perspectives of Caesar's works.

Based on the assumption that Latin texts written by the same author are more similar to each other than to those by other authors, we have two specific hypotheses. (1) All three chapters of the *Civil War* and Chapters 1–7 of the *Gallic War* are more similar to each other than to Chapter 8 of the *Gallic War*. (2) The *African War*, *Alexandrine War*, and *Spanish War* are dissimilar to each other and dissimilar to the *Civil War* and the *Gallic War* (Ch. 1–7).

Methods

Materials

The Latin texts of Caesar's war memoirs were downloaded from the online Latin Library[2]. All words and numbers that do not belong to the core writings were removed (e.g., page numbers). *Commentarii de Bello Civili* (*Civil War*) has three chapters that describe the events of the Great Roman Civil War (49–45 BC), all written by Caesar. *Commentarii de Bello Gallico* (*Gallic War*) has eight chapters about Caesar's campaigns in Gaul and southern Britain in the 50s BC. He wrote the first seven chapters; the last chapter was written after his death by one of his lieutenants, Aulus Hirtius (Carter 1997). *De Bello Alexandrino* (*Alexandrine War*), *De Bello Africo* (*African War*), and *De Bello Hispaniensi* (*Spanish War*) are about Caesar's campaigns in Alexandria, North Africa, and the Iberian Peninsula. The authorship of these three commentaries has often been attributed to Caesar, but has been under much debate (Carter 1997; Conte 1994; Hall 1996; Storch 1977).

Distributional Semantics

Methods of distributional semantics (reviewed in Cohen & Widdows 2009 and Turney & Pantel 2010) are approaches to automated analysis of natural language text ("natural" refers to human languages, as opposed to artificial computer languages). They have wide applications in automated

interpretation of written text and have been successfully used to study the authorship of certain books of the New Testament (Widdows & Cohen 2009). For authorship attribution, the key assumption is that texts written by the same author should be more similar to each other than to those written by other authors. Geometric methods of distributional semantics, such as Latent Semantic Analysis, or LSA, (Deerwester *et al.* 1990), derive vector representations of words from the contexts in which they occur (such as within documents, as in the current study). The fundamental vector representations in the current study are components of a Term-by-Document matrix (more specifically, an approximation of this matrix after dimension reduction), with each column vector representing a different document, each row vector representing a different word, and each cell representing the frequency with which a word occurs in a particular document, or some statistical transformation of this value. For example, in the TermDocument matrix in Table 1, the raw frequencies of the Latin words *unaque* and *una* from Chapters 1, 2, and 8 of *Gallic War* were counted.

Original TermDocument Matrix (raw frequencies)

		Document Vectors		
		Gallic1	Gallic2	Gallic3
Term Vectors	unaque (one)	0	0	3
	una (one)	9	11	0
	more words	…	…	…

Table 1. TermDocument matrix counting the raw frequencies of the Latin words unaque *and* una *from Chapters 1, 2, and 8 of the* Gallic War.

Each row is a vector of the frequencies with which a particular word appears in every document (columns). Vector representations of larger units of text, such as documents, are generated by adding these term vectors together. This facilitates the estimation of semantic relatedness

between textual units – for example, the cosine similarity (normalized scalar product) between document vectors after dimension reduction can be used to measure the similarity between two documents (Landauer, Foltz & Laham 1998). The resulting document-to-document similarity metrics can then be utilized in downstream tasks, such as text categorization (Vasuki & Cohen 2010) and automated essay grading (Landeaur *et al.* 1997). *Random Indexing* (RI) (Kanerva, Kristofersson & Holst 2000) is a relatively recent distributional semantics method. RI uses a form of random projection (for a detailed account of random projection, see Vempala 2005) to create an approximation of a Term-by-Context matrix, without the need to represent the full matrix in its entirety. In this way, RI addresses the computational challenges imposed by other dimension reduction methods such as the Singular Value Decomposition (SVD, see Golub & Reinsch 1970), and it has been shown to reduce the dimensionalities of such matrices without compromising performance of the resulting models on cognitive tasks (Kanerva, Kristofersson & Holst 2000). The net result is a Term-by-*k* matrix, with *k* as the user-defined dimensionality of the space. *k*-dimensional document vectors are generated by adding together the term vectors for terms occurring in a document. Finally, the cosine similarity between each pair of documents is calculated.

Software Package

The RI implementation used in this project is provided by the SemanticVectors package version 5.4 (Widdows & Ferraro 2008; Widdows & Cohen 2010), an open source software package originally developed as part of a project initiated at the University of Pittsburgh[3]. The software was run on a 13' MacBook Pro with the Java SE Development Kit 7 for Mac OS X x64. All parameters such as seedlength (default = 10; the number of stochastically assigned non-zero values for the vectors used to initialize the RI procedure), minfrequency and maxfrequency (default = no constraints; the frequency boundaries within which a word must

occur in the corpus to be considered), term weight (default = none; no statistical transformations applied to raw occurrence frequencies), and vectortype (default = real vectors; determining whether the underlying vector space is real, binary or complex in nature) were retained as their default values, except that the *k* parameter (default = 200) for dimension reduction was set to 500.

Design

The design is based on the assumption that aspects of a particular author's writing style are statistically measurable, relatively consistent across his/her works, and reliably different in other authors' works. The *independent variables* are (1) the authors (Caesar, Hirtius, or other) and (2) the documents (chapter or whole book). The *dependent variable* is the cosine similarity score between each pair of document vector representations of the text constructed by the Semantic Vectors package. Cosine Similarity between Vector A and Vector B is defined by:

$$\text{Similarity} = \text{Cos}(\theta) = \frac{A \bullet B}{\|A\| \|B\|} = \frac{\sum_{i=1}^{n} A_i \times B_i}{\sqrt{\sum_{i=1}^{n} (A_i)^2 \times \sum_{i=1}^{n} (B_i)^2}}$$

Procedure

Following the procedure described in the previous section, RI was performed to generate 500-dimensional vectors for each text of interest (which is to say, the initial representation was an approximation of a term-by-text matrix, and the rows of this matrix – the word vectors – were added together to generate vector representations of texts). Then each pair of texts was compared, resulting in a set of pair-wise cosine similarity scores. The texts were chapters when chapters were compared,

and books when entire books were compared. The following are the details of the steps using version 5.4 of the SemanticVectors software.

1. The original TermDocument matrix was created for each set of documents to be compared.
   - java pitt.search.lucene.IndexFilePositions TextFile (e.g., Caesar)
2. A reduced TermDocument matrix with 500 dimensions using Random Index algorithm was created:
   - java pitt.search.semanticvectors.BuildIndex -luceneindexpath positional_index -dimension 500
3. Each pair of documents were compared:
   - java pitt.search.semanticvectors.CompareTerms -queryvectorfile docvectors.bin ChapterX ChapterY (e.g., Gallic_Chapter_1 Gallic_Chapter_8)
4. Each pair of books were compared:
   - java pitt.search.semanticvectors.CompareTerms -queryvectorfile docvectors.bin Book1 Book2 (e.g., Civil_War African_War)

To visualize the differences between texts, we projected them into a two-dimensional space using the SemanticVectors utility for this purpose, which applies Principal Component Analysis to groups of high-dimensional vector representations.

To further validate our approach we performed four additional studies. First, we selected a 3-section work, *Quintum Fratrem,* by Marcus Tullius Cicero that is comparable in length to Caesar's 3-chapter *Civil War* (word counts of 7 933, 5 504 and 5 496 for *Quintum Fratrem* vs. word counts of 10 998, 6 433, and 15 151 for *Civil War*). Cicero was a notable contemporary of Caesar and a prolific writer, orator, and philosopher in the ancient Roman Republic. These two books are known to be

written by Cicero and Caesar. We used the same method to compare the chapters. By comparable lengths we mean that the lengths of texts are in the same order of magnitude, not in equal numbers of word count. We were not able to obtain texts that are equal or very close to each other in length, as the texts are historical documents that are limited in number. We did not attempt to normalize these document lengths (for example by subsampling terms), as the cosine comparison between document vectors considers the relative orientation of these vectors (the distribution of terms they contain), rather than their lengths (the counts of terms they contain).

Second, to further verify that the estimates provided by this method are sufficiently topic independent to be attributable to authorship, we picked three works of three different topics by Marcus Tullius Cicero: 1) *In Catilinam*, legal prosecution speeches that Cicero used to expose a major conspiracy, 2) *De Officiis*, a work on ethics outlining Cicero's view of the best way to live life and observe moral obligations, and 3) *Quintum Fratrem*, a series of personal letters Cicero wrote to his brother. These three works and Caesar's work the *Civil War* (observably, a book about war) were compared with each other using the same distributional semantics method as the initial study.

Third, we evaluated the extent to which stochastic initialization in RI affected our results. RI is not deterministic and its repeated application to the same corpus may produce different results. Consequently, we generated and compared the representations of the *Civil War* and *Gallic War*, and *Gallic War* and *African War* 10 times, with different random vectors initializing each iteration.

Finally, we evaluated the influence of the dimensionality parameter on the performance of our models. We re-ran two comparisons (*Civil War* vs. *Gallic War* and *Gallic War* vs. *African War*) with dimensions ranging from 10 to 100 with increments of 10 (a total of 10 comparisons), and

from 200 to 1100 with increments of 100 (also a total of 10 comparisons).

Results

War Memoirs by Caesar and War Memoirs Attributed to Caesar

The similarity scores between each pair of Gallic chapters are shown in Table 2. Larger numbers reflect higher similarities, with a maximum possible cosine similarity of 1.0 for identical document vectors.

	Gallic1	*Gallic2*	*Gallic3*	*Gallic4*	*Gallic5*	*Gallic6*	*Gallic7*	*Gallic8*	Mean
Gallic1		0.8948	0.8642	0.8852	0.8677	0.8541	0.8712	**0.8380**	0.8679
Gallic2	0.8948		0.8841	0.9003	0.8901	0.8839	0.8804	**0.8600**	0.8848
Gallic3	0.8642	0.8841		0.8951	0.8810	0.8618	0.8762	**0.8456**	0.8726
Gallic4	0.8852	0.9003	0.8951		0.8936	0.8807	0.8779	**0.8529**	0.8837
Gallic5	0.8677	0.8901	0.8810	0.8936		0.8893	0.8821	**0.8513**	0.8793
Gallic6	0.8541	0.8839	0.8618	0.8807	0.8893		0.8699	**0.8357**	0.8679
Gallic7	0.8712	0.8804	0.8762	0.8779	0.8821	0.8699		**0.8554**	0.8733
Gallic8	**0.8380**	**0.8600**	**0.8456**	**0.8529**	**0.8513**	**0.8357**	**0.8554**		**0.8484**
Mean	0.8679	0.8848	0.8726	0.8837	0.8793	0.8679	0.8733	**0.8484**	0.8722

Table 2. Chapter by Chapter Comparisons for the Gallic War.

As can be seen in Table 2, Chapters 1 to 7 of the *Gallic War* are more similar to each other than to Chapter 8. A statistical analysis was performed, using an unpaired two-tail *t*-test. For each chapter (e.g., Gallic1) its similarities to all other chapters were examined. Then the *t*-test was used to evaluate the hypothesis that the average similarity between a particular chapter and all others would be different from the average similarities amongst these other chapters if the evaluation were run on a much larger set of documents (see Table 3). For example, to compare the similarity between Gallic1 and Gallic8, the similarities between Gallic1 vs. {Gallic2 through Gallic7} were compared with the similarities between

Gallic8 vs. {Gallic1 through Gallic 7}, using a two-tailed *t*-test. As seen in Table 3, which shows the *p*-values produced by the pairwise *t*-tests across chapters, only the differences between the other seven chapters and Gallic8 are statistically significant (smallest $t(12) = 2.46$, with largest $p < 0.03$). That is, the similarity between RI representation of Chapter 8 of the *Gallic War* and the RI representation of the other chapters is significantly different from the average similarity between the RI representations of the other seven chapters. This result is a confirmation of the known fact that Chapter 8 of the *Gallic War* was written by a different author after Caesar's death.

	Gallic1	Gallic2	Gallic3	Gallic4	Gallic5	Gallic6	Gallic7	Gallic8
Gallic1		0.0740	0.6298	0.1152	0.2339	0.9978	0.5065	**0.0298**
Gallic2			0.1491	0.8862	0.4767	0.0750	0.0782	**0.0001**
Gallic3				0.2231	0.4406	0.6325	0.9202	**0.0053**
Gallic4					0.6052	0.1165	0.1578	**0.0003**
Gallic5						0.2359	0.3830	**0.0005**
Gallic6							0.5098	**0.0300**
Gallic7								**0.0002**
Gallic8								

Table 3. p-*values of* t-*tests for* Gallic War *Chapter Comparisons. Boldface indicates* $p < 0.05$.

For similarities across the five war memoirs, the similarity score (0.9126) between the *Civil War* and Caesar's contribution to the *Gallic War* (the first seven chapters) is compared with the average similarity between each of the *Civil War* and *Gallic War* and each of the *African War*, *Spanish War*, and *Alexandrine War* (the other six scores in the first two rows of Table 4; average similarity = 0.8886). The similarity between the *Civil War* and *Gallic War* (both confirmed to be written by Caesar) is significantly different from their similarity to the other three war memoirs ($t(5) = 5.68$, $p = 0.002$), confirming that the *Civil War* and *Gallic War* are more similar to each other than to the other three war memoirs.

	Gallic	Civil	African	Spanish	Alexandrine
Gallic		**0.9126**	0.8701	0.8902	0.8921
Civil			0.887	0.8902	0.9018
African				0.8681	0.8597
Spanish					0.8654
Alexandrine					

Table 4. Similarity Score Matrix of Civil War *&* Gallic War *vs.* African War, Alexandrine War, *&* Spanish War.

These metrics of similarity do not provide the means to draw conclusions about a single pair of books. Rather, they provide the means to detect outliers – volumes that are significantly different from the rest of a collection. Our studies suggest that a volume with similarity to the rest of a collection that is statistically significantly lower than the mean similarity between the remaining volumes may have a different author.

To further analyze the similarities across the documents, we visualized the similarities by generating a reduced-dimensional approximation of the document vectors using Singular Value Decomposition (Widdows & Cederberg 2003). The second and third dimensions of the reduced-dimensional space were used as x and y coordinates (Fig. 1).

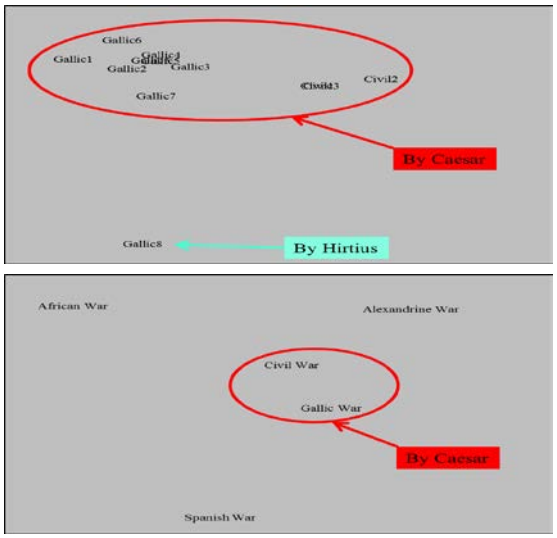

Figure 1. Two-dimensional production of the RI representations of the chapters of Civil War *and* Gallic War *(top panel) and the five war memoirs.*

Displaying the second and third dimension is the default in the Semantic Vectors package, as the first dimension tends not to be informative when visualizing semantic vectors in this manner (for further details on this point, see Widdows 2004). Shorter distances between documents indicate higher similarities. When the three chapters of the *Civil War* and the eight chapters of the *Gallic War* were compared with each other, Chapter 8 of the *Gallic War* stood out by itself. In addition, the *Civil War* and *Gallic War* are similar to each other but are different from the *African War*, *Alexandrine War*, and *Spanish War*, which are also different from each other (see Fig. 1). This result suggests that the latter three were

not authored by Caesar; furthermore, they might be written by multiple authors.

Caesar vs. Cicero

The average similarity is 0.8027 between Cicero's three chapters and 0.8497 between Caesar's three chapters (μ=0.8262 for within-author similarity) but μ=0.6457 for across-author comparison, i.e., across each pair of Cicero's and Caesar's chapters. The average similarity for within-author comparisons and the average similarity for across-author comparisons are significantly different (μ=0.8262 vs. μ=0.6457, $t(13) = 17.91$, $p < 0.0001$). This clearly shows that the same distributional semantics method applied equally well in the disambiguation of known authorship between Caesar and Cicero, and this further supports that the estimated difference between Chapter 8 of the *Gallic War* and the other chapters by Caesar is related to aspects of authorship.

Topics vs. Authorship

Table 5 shows the results of the comparisons among Cicero's three books with different topics and Caesar's *Civil War*.

	Caesar: *Civil War* (32 582 words)	Cicero: *In Catilinam* (12 745 words)	Cicero: *De Officiis* (35 001 words)	Cicero: *Quintum Fratrem* (18 933 words)
Caesar: *Civil War*		**0.7524**	**0.6923**	**0.7271**
Cicero: *In Catilinam*			0.8133	0.8222
Cicero: *De Officiis*				0.8188
Cicero: *Quintum Fratrem*				

Table 5: Similarity between Books of Different Topics.

Results show that while the four books concerned quite different topics, the *Civil War* by Caesar is clearly less similar to each of Cicero's works

than Cicero's works are with each other, as shown by the similarity scores in the table below (μ = 0.7239 vs. μ = 0.8181, $t(4) = 5.35$, $p = 0.006$). While the topics differed, Cicero's works were relatively more similar to other works of his than they were to Caesar's book, and this similarity did not vary much across pairs of books that Cicero authored. Thus, the topics of the texts do not significantly affect the similarity scores.

Stochastic Initialization

To show that random seeds in our study did not affect the results, we regenerated our distributional models and repeated our comparisons between the *Civil War* and *Gallic War* and the *Gallic War* and *African War* 10 times, each run with a different random seed. All runs yielded nearly identical results: for *Civil War* vs. *Gallic War*, mean similarity μ=0.9138, SD = 0.0008, Relative SD = 0.09%; for *Gallic War* vs. *African War*, μ=0.8821, SD = 0.0049, and Relative SD = 0.56%. The very small Relative Standard Deviation is an indication that the initial random seeds did not have a significant impact on the results.

Dimensionality

When the number of dimensions is small (from 10 to 100), the results fluctuate noticeably. For *Civil War* vs. *Gallic War* comparisons, Min = 0.8615, Max = 0.9428, SD = 0.0219, Relative SD = 2.4%, and for *Gallic War* vs. *African War*, Min = 0.8675, Max = 0.9179, SD = 0.0159, and Relative SD = 1.78%. When the number of dimensions is larger (from 200 to 1100), the results are almost identical. For *Civil War* vs. *Gallic War* comparisons, Min = 0.9124, Max = 0.9140, SD = 0.0006, Relative SD = 0.07%, and for *Gallic War* vs. *African War*, Min = 0.8829, Max = 0.8857, SD = 0.0008, and Relative SD = 0.09%. The results fluctuate very little after 200 dimensions, which provides evidence that a dimensionality of 500 is adequate for our current purposes, as a Relative Standard Deviation of <0.1% is sufficiently small and does not make a sizable impact on the measurement.

Discussion & Conclusion

The current computational study of Classical Latin authorship using distributional semantics is consistent with the known fact that Chapter 8 of the *Gallic War* was not written by Caesar. This provides support for the utility of RI as a method for authorship attribution of Classical Latin text. It further suggests that the *African War*, *Alexandrine War*, and *Spanish War* were unlikely to have been written by Caesar, providing fresh evidence toward resolving the two-millennium old mystery of Caesar's authorship of the war memoirs.

In classical studies, authorship attribution is typically based on literary analysis. In the analysis of the authorship of the works attributed to Caesar, classical scholars analyzed the contents, writing styles, and author perspectives of the Latin texts (see Carter 1997, for a review; see also Conte 1994; Hall 1996; Storch 1977). For example, Caesar's narratives in *Civil War* and *Gallic War* are designed to justify his actions and his victories and include strategic materials that were only available to him, whereas the writings in other works attributed to Caesar have a narrower perspective and do not appear to have a driving agenda; they merely chronicle the details of events such as the details of camp life, fighting and training. This type of literary analysis was used to suggest that Caesar did not write *Spanish War* (Carter 1997). However, such literary analysis methods are often subjective. Our current study provides an alternative to literary analysis for authorship attribution of Latin texts. Distributional semantics is an automated process that can be easily scaled up to process large amounts of Latin texts that would be prohibitively time-consuming to analyze manually. In addition, it is based on objective measures.

Topic models for authorship attribution, such as those explored by Seroussi (2014) and Savoy (2013) based on LDA (Latent Dirichlet Allocation, Blei 2012), also belong to the family of methods of distributional semantics. They provide a different and arguably complementary approach to the broader problem of authorship attribution. As the name

suggests, topic models are models of the thematic content of text. The goal of constructing a topic model is to identify a set of latent topics that characterize a set of documents. Topics are modeled as distributions over terms, which is to say the probability of seeing a term when a particular topic is discussed is inferred during the construction of the model. Documents, in turn, are modeled as distributions over topics that can then be leveraged for supervised machine learning, as documents discussing similar subjects will have similar probability distributions over the inferred topics. Topic model features of this sort have been shown to be useful as a means to attribute authorship across relatively large corpora produced by multiple authors, and spanning multiple topics (Seroussi 2014; Savoy 2013), although it is worth noting that frequently-used function words predominate in some of the inferred topics in this prior work. In contrast, the main purpose of the current work using distributional semantics is to identify outlier authorship within the context of a small set of chapters with similar topics.

Although this study concerns the application of a specific computational method to a specific authorship attribution problem involving Caesar's writings, it illustrates the broader potential of modern computational methods as a means to conduct classical studies. Classical scholars and computational linguists have started to work together to build digital infrastructures such as digital repositories and treebanks for the Latin language and to develop computational algorithms to analyze Latin texts (e.g., Bamman & Crane 2007, 2009; Forsall *et al*. 2014; McGillivray 2013; Scheirer, Forstall & Coffee 2016). For example, Forstall *et al*. (2014) used word-level n-gram matching to study intertextuality of Latin texts and showed that lemma identity, word frequency, and phrase density are important constituents of what make a phrase parallel a meaningful intertext. With more ancient Latin texts being digitized, computational methods can be developed to study and answer many significant questions about the evolution of the Latin language, the change of culture and politics as reflected in written texts, digital "fingerprinting" of historical

documents, and so on. This trend is for all languages, not just for Latin. These recent developments have contributed to the emergence of a new field – *digital humanities*.

While the current study is focused on ancient Latin literature, it also has significant implications for contemporary problems. For example, similar distributional semantics techniques could be used to develop software applications to detect fake reviews in online shopping and online services, to discover plagiarism, to identify the gender, age, and native language of the author, and so on.

Unlike other statistical and computational methods that focus on local properties and units at lower levels (see Mendenhall 1887 and Mascol 1888a/b; Mosteller and Wallace 1964; Juola 2008) such as frequencies of pre-identified words or phrases, or frequencies of function words (e.g., pronouns, determiners, and prepositions in contrast with content words such as nouns and verbs), RI does not require manual selection of such features, which are typically not yet identified in languages other than English. In fact, some of these features emerge automatically as a result of the analysis. For example, and as shown in Table 6, the nearest terms (words) to a chapter vector for the *Gallic War* study are automatically generated and the top ten such terms for all the chapters compared in the *Gallic War* study include 70 percent function words and pronouns which are context free or relatively independent from the content words (verbs, nouns, adjectives, etc.). So these function words have come to dominate the representations of texts, presumably on account of their high relative frequency. These results suggest an important constraint on the application of our method is that the chapters considered be of sufficient length for frequently occurring terms to be representationally dominant. While this appears to be the case for all of the works used in the current proposal, the nature of the relationship between document length and the representational predominance of function words remains an open question to pursue in future work.

RANK	Gallic 1	Gallic 2	Gallic 3	Gallic 4
1	***non*** not	***qui*** who	***ac*** and	***neque*** or
2	***quod*** that	***inter*** between	*magno* great	*capere* to catch
3	***si*** if	***eorum*** those	*fere* almost	*partem* part
4	***esse*** to be	***ad*** to	***neque*** or	***multis*** many
5	***sibi*** himself	***in*** in	***in*** in	*proelio* battle
6	***quam*** than	*tanta* great	***ad*** to	*reliqui* the rest
7	***minus*** less	*tempus* time	***quod*** that	***se*** himself
8	***id*** id	***et*** and	***et*** and	***ad*** to
9	***eam*** her	*multitudine* multitude	*eadem* the same	***et*** and
10	***ea*** it	***quod*** that	*ex* from	***quibus*** which
Function Words	10	7	6	6

RANK	Gallic 5	Gallic 6	Gallic 7	Gallic 8
1	***tum*** then	*numero* number	***omni*** all	*civitates* cities
2	***nihil*** nothing	*eadem* the same	*animo* mind	***cum*** with
3	***re*** re	***aut*** or	***ne*** do not	***quorum*** their
4	***omnibus*** all	*bello* war	***qua*** which	***esset*** was
5	***pro*** for	*una* one	***ab*** from	***autem*** however
6	***ad*** to	***reliquis*** other	***parte*** part	*undique* round
7	***circiter*** about	*tribus* tribe	***iam*** already	***sine*** without
8	*vallo* rampart	*magno* great	*relictis* leaving	***contra*** against
9	*prima* first	***in*** in	***quid*** what	***ipse*** himself
10	***partem*** part	***fere*** almost	***omnibus*** all	*legiones* legions
Function Words	8	4	8	7

Table 6. Nearest neighboring terms to the chapter vector representations for Gallic War.

There are several limitations in the current study. For example, the sample size is small: only the five war memoirs attributed to Caesar were studied initially. We later added a comparison of Caesar's 3-chapter *Civil War* with Cicero's 3-chapter *Quintum Fratrem*, which are comparable in lengths and in historical context. We also made a comparison between these two works as well as two other works of Cicero, all on different topics, to confirm that this method yielded topic independent results. Though these additional studies provide support for the validity of our approach, further studies with a larger corpus of Latin including texts by other authors are needed to establish the constraints on its generalizability. Another limitation is the assumption that an author's writing style does not change significantly over time. If this is not the case, then the date of authorship of writings should to be considered in the comparisons. These issues are being addressed in an ongoing follow-up study, in which we use distributional semantics and other computational methods to understand the evolution of the Latin language over the past two millennia.

Olivia Zhang is a high school senior at St. John's School in Houston, Texas, USA. Her primary research interests concern computational linguistics and the applications of technology to diverse fields in the humanities, which she plans to pursue at Harvard University in her undergraduate studies.

Contact: oliviarzhang8@gmail.com

Dr. Scott McGill is a professor and Department Chair of Classical and European Studies at Rice University in Houston, Texas, USA. His research interests include Latin poetry in late antiquity, Virgil and Virgil's reception, and Roman literary culture. He is the author of three books and has edited three collections of essays.

Contact: smcgill@rice.edu

Dr. Trevor Cohen is an associate professor in the School of Biomedical Informatics at the University of Texas Health Science Center at Houston (UTHealth), USA. His research focuses on the application of methods of distributional semantics to biomedical problems, with applications in literature-based discovery, post-marketing surveillance of pharmaceuticals, cancer-related drug repurposing and information retrieval.

Contact: trevor.cohen@uth.tmc.edu

Notes

1. It should be noted that automated authorship attribution cannot be conducted without providing some form of information to an automated system, and that the words in the chapters we analyze in the current study can be further decomposed into characters, or pre-processed with a parser or part-of-speech tagger to enrich it with syntactic features. When we refer to feature engineering, we are referring to the imposition of additional constraints on the "raw" information provided, based on preconceptions about what components of this information would be most helpful for the classification task at hand. For example, we consider manual selection of a subset of terms in a vocabulary based on the hypothesis that these specific terms would be more useful for classification to be feature engineering, just as we would view the imposition of, for example, a Gabor filter for edge detection on a gray-scale image. Arguably, part-of-speech tagging would also be a form of feature engineering. However, the lexical features used for modeling in distributional semantics are generally neither manually selected, nor added to the raw data through a pre-processing procedure. Thus, we would argue that feature engineering is not a requirement of our method.
2. At http://www.thelatinlibrary.com
3. The package is available at https://github.com/semanticvectors/semanticvectors/

References

ARGAMON, SHLOMO ET AL. (2003). "Gender, Genre, and Writing Style in Formal Written Texts." *Text* 23.3: 321–346.

ALTHENEYAN, ALAA SALEH & MOHAMED EL BACHIR MENAI (2014). "Naïve Bayes Classifiers for Authorship Attribution of Arabic Texts." *J. King Saud Univ. Comput. Inf. Sci.* 26.4: 473–484.

BAAYEN, HARALD ET AL. (2002). "An Experiment in Authorship Attribution." *Proceedings of JADT 2002, Universite de Rennes, St. Malo*. 29–37.

BAMMAN, DAVID & GREGORY CRANE (2007). "The Latin Dependency Treebank in a Cultural Heritage Digital Library." *Proceedings of the Workshop on Language Technology for Cultural Heritage Data (LaTeCH 2007), Prague, 28 June 2007.* Association for Computational Lingustics. 33–40.

BAMMAN, DAVID & GREGORY CRANE (2009). "Computational Linguistics and Classical Lexicography." *Digital Humanities Quarterly*, 3.1.

BLEI, DAVID M. (2012). "Probabilistic Topic Models." *Communication of the ACM* 55: 77–84.

BURGER, JOHN D. & JOHN C. HENDERSON (2006). "An Exploration of Observable Features Related to Blogger Age." *Proceedings of AAAI Spring Symposium on Computational Approaches to Analyzing Weblogs.* New York, NY: ACM Press. 47–54.

BURROWS, JOHN F. (2002). "Delta: A Measure of Stylistic Difference and a Guide to Likely Authorship." *Literary and Linguistic Computing* 17: 267–287.

CARTER, JOHN (1997). *Julius Caesar – The Civil War with Anonymous Alexandrian, African, and Spanish Wars* (Translated with an Introduction and Notes by John Carter). Oxford, UK: Oxford University Press.

COHEN, TREVOR & DOMINIC WIDDOWS (2009). "Empirical Distributional Semantics: Methods and Biomedical Applications." *Journal of Biomedical Informatics* 42.2: 390–405.

CONTE, GIAN BIAGIO (1994). *Latin Literature: A History* (Translated by Solodow, J. B). Baltimore: Johns Hopkins University Press.

DEERWESTER, SCOTT ET AL. (1990). "Indexing by Latent Semantic Analysis." *Journal of the American Society for Information Science* 41: 391–407.

DUMAIS, SUSAN T. (2003). "Data-driven Approaches to Information Access." *Cognitive Science* 27: 491–524.

FORSTALL, CHRISTOPHER ET AL. (2014). "Modeling the Scholars: Detecting Intertextuality through Enhanced Work-level n-gram Matching." *Digital Scholarship in the Humanities* 30.4: 503–515.

FOX, NEAL, OMRAN EHMODA & EUGENE CHARNIAK (2012). "Statistical Stylometrics and the Marlowe–Shakespeare Authorship Debate." *Proceedings of the Georgetown University Roundtable on Language and Linguistics (GURT), Washington, D.C, USA*.

GOLUB, GENE H. & C. REINSCH (1970). "Singular Value Decomposition and Least Squares Solutions." *Numerische Mathematik* 14: 403–20.

GRAHAM, NEAL, GRAEME HIRST & BHASKARA MARTHI (2005). "Segmenting Documents by Stylistic Character." *Natural Language Engineering* 11:4: 397–415.

GRIEVE, JACK (2007). "Quantitative Authorship Attribution: An Evaluation of Techniques." *Literary and Linguistic Computing* 22: 251–270.

HALL, LINDSAY G. H. (1996). "Hirtius and the Bellum Alexandrinum." *Classical Quarterly 46*.2: 411–415.

HOLMES, T. RICE (2011). *The Roman Republic and the Founder of the Empire*. BibliLife.

JUOLA, PATRICK (1997). "What Can We Do With Small Corpora? Document Categorization via Cross-Entropy." *Proceedings of an Interdisciplinary Workshop on Similarity and Categorization, Department of Artificial Intelligence, University of Edinburgh, Edinburgh, UK.*

JUOLA, PATRICK (2007). "Becoming Jack London." *Journal of Quantitative Linguistics* 14.2: 145–147.

JUOLA, PATRICK (2008). "Authorship Attribution." *Foundations and Trends in Information Retrieval* 1.3: 233–334.

KANERVA, PENTTI, JAN KRISTOFERSSON & ANDERS HOLST (2000). "Random Indexing of Text Samples for Latent Semantic Analysis." *Proceedings of 22nd Conference of Cognitive Science Society*. 1036.

KOPPEL, MOSHE, SHLOMO ARGAMON & RACHEL SHIMONI (2002). "Automatically Categorizing Written Texts by Author Gender." *Literary and Linguistic Computing* 17.4: 401–412.

KOPPEL, MOSHE, JONATHAN SCHLER & SHLOMO ARGAMON (2009). "Computational Methods in Authorship Attribution." *Journal of the American Society for Information Science and Technology* 60.1: 9–26.

KOPPEL, MOSHE, JONATHAN SCHLER & KFIR ZIGDON (2005). "Determining an Author's Native Language by Mining a Text for Errors." *KDD '05 Proceedings of the Eleventh ACM SIGKDD International Conference on Knowledge Discovery in Data Mining, Chicago, IL*. 624–628.

LANDAUER, THOMAS K., PETER W. FOLTZ & DARRELL LAHAM (1998). "An Introduction to Latent Semantic Analysis." *Discourse Processes* 25: 259–284.

LANDAUER THOMAS K. ET AL. (1997). "How Well Can Passage Meaning be Derived without Using Word Order? A Comparison of Latent Semantic Analysis and Humans." *Proceedings of the Nineteenth Annual Conference of the Cognitive Science Society, August 7–10, 1997, Stanford University.*

MCGILLIVRAY, BARBARA (2013). *Methods in Latin Computational Linguistics*. Leiden, Netherlands: Brill Academic Publishers.

MASCOL, CONRAD (1888a). "Curves of Pauline and Pseudo-Pauline Style I." *Unitarian Review* 30: 452– 460.

MASCOL, CONRAD (1888b). "Curves of Pauline and Pseudo-Pauline Style II." *Unitarian Review* 30: 539–546.

MATTHEWS, ROBERT A. J. & THOMAS V. N. MERRIAM (1993). "Neural Computation in Stylometry : An Application to the Works of Shakespeare and Fletcher." *Literary and Linguistic Computing* 8.4: 203–209.

MENDENHALL, T. C. (1887). "The Characteristic Curves of Composition." *Science* 11: 237–249.

MERRIAM, THOMAS V. N. & ROBERT A. J. MATTHEWS (1994). "Neural Computation in Stylometry II: An Application to the Works of Shakespeare and Marlowe." *Literary and Linguistic Computing 9*: 1–6.

MITCHELL, TOM M. (1997). *Machine Learning*. New York: McGraw-Hill.

MOSTELLER, FREDERICK & DAVID L. WALLACE (1963). "Inference in Authorship Problem – A Comparative Study of Discrimination Methods Applied to Authorship of Disputed Federalist Papers." *J. Am. Stat. Assoc.* 58: 275–309.

PENG, FUCHUN, DALE SCHUURMANS & SHAOJUN WANG (2004). "Augmented Naive Bayes Text Classifier with Statistical Language Models." *Information Retrieval* 7.3–4: 317–345.

SAVOY, JACQUES (2013). "Authorship Attribution Based on a Probabilistic Topic Model." *Information Processing & Management* 49.1: 341–354.

SCHEIRER, WALTER, CHRISTOPHER FORSTALL & NEIL COFFEE (2016). "The Sense of a Connection: Automatic Tracing of Intertextuality by Meaning." *Digital Scholarship in Humanities* 31.1: 204–217.

SCHLER, JONATHAN ET AL. (2006). "Effects of Age and Gender on Blogging." *Proceedings of the AAAI Spring Symposium on Computational Approaches for Analyzing Weblogs.* Menlo Park, CA: AAAI Press. 199–205.

SCHUTZE HINRICH (1998). "Automatic Word Sense Discrimination." *Computational Linguistics* 24: 97–123.

SEROUSSI, YANIR, INGRID ZUKERMAN & FABIAN BOHNERT (2014). "Authorship Attribution with Topic Models." *Computational Linguistics* 40.2: 269–310.

STAMATATOS, EFSTATHIOS (2009). "A Survey of Modern Authorship Attribution Methods." *Journal of the American Society for Information Science and Technology* 60.3: 538–556.

STORCH, RUDOLPH H. (1977). "The Author of the deBello Hispaniensi: A Cavalry Officer?" *Acta Classica (Cap Town)* 20: 201–204.

TURNEY, PETER & PATRICK PANTEL (2010). "From Frequency to Meaning: Vector Space Models of Semantics." *Journal of Artificial Intelligence Research* 37: 141–188.

ULE, LOUIS (1982). "Recent Progress in Computer Methods of Authorship Determination." *Association for Literary and Linguistic Computing Bulletin* 10: 73–89.

VAN HALTEREN, HANS ET AL. (2005). "New Machine Learning Methods Demonstrate the Existence of a Human Stylome." *Journal of Quantitative Linguistics* 12.1: 65–77.

VASUKI, VIDYA & TREVOR COHEN (2010). "Reflective Random Indexing for Semi-Automatic Indexing of the Biomedical Literature." *Journal of Biomedical Informatics* 43.5: 694–700.

VEMPALA, SANTOSH S. (2005). *The Random Projection Method*. American Mathematical Society.

WIDDOWS, DOMINIC (2004). *Geometry and Meaning*. CSLI Publications.

WIDDOWS, DOMINIC & TREVOR COHEN (2009). "Semantic Vector Combinations and the Synoptic Gospels." *Quantum Interaction: Lecture Notes in Computer Science* 5494: 251–265.

WIDDOWS, DOMINIC & TREVOR COHEN (2010). "The Semantic Vectors Package: New Algorithms and Public Tools for Distributional Semantics." *Fourth IEEE International Conference on Semantic Computing (IEEE ICSC2010), Pittsburgh, September 22–24, 2010*.

WIDDOWS, DOMINIC & SCOTT CEDERBERG (2003). "Monolingual and Bilingual Concept Visualization from Corpora." *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology: Demonstrations,* 4: 31–32.

WIDDOWS, DOMINIC & KATHLEEN FERRARO (2008). "Semantic Vectors: A Scalable Open Source Package and Online Technology Management Application." *Proceedings of Sixth International Conference on Language Resources and Evaluation (LREC 2008)*.

ZHAO, YING & JUSTIN ZOBEL (2005). "Effective Authorship Attribution Using Function Word." *Proc. 2nd AIRS Asian Information Retrieval Symposium*, Springer. 174–190.