

# Large-scale quantitative profiling of the Old English verse tradition

Leonard Neidorf<sup>1,7</sup>, Madison S. Krieger<sup>2,7\*</sup>, Michelle Yakubek<sup>3,4</sup>, Pramit Chaudhuri<sup>5</sup> and Joseph P. Dexter<sup>6\*</sup>

**The corpus of Old English verse is an indispensable source for scholars of the Indo-European tradition, early Germanic culture and English literary history. Although it has been the focus of sustained literary scholarship for over two centuries, Old English poetry has not been subjected to corpus-wide computational profiling, in part because of the sparseness and extreme fragmentation of the surviving material. Here we report a detailed quantitative analysis of the whole corpus that considers a broad range of features reflective of sound, metre and diction. This integrated examination of fine-grained features enabled us to identify salient stylistic patterns, despite the inherent limitations of the corpus. In particular, we provide quantitative evidence consistent with the unitary authorship of *Beowulf* and the Cynewulfian authorship of *Andreas*, shedding light on two longstanding questions in Old English philology. Our results demonstrate the usefulness of high-dimensional stylometric profiling for fragmentary literary traditions and lay the foundation for future studies of the cultural evolution of English literature.**

Composed between roughly 600 and 1100, Old English literature represents the earliest phase of literary production in English. Although it is assumed that most works of Old English literature have not survived, the remainder nevertheless encompass not only a broad time period but also multiple streams of influence—Germanic, Christian, and classical Greek and Roman—as well as diverse genres such as heroic poetry, riddles and biblical works<sup>1</sup>. This rich corpus also contains one of the masterpieces of English literature—the epic poem *Beowulf*. Both for its historical importance and its aesthetic merit, Old English literature has attracted the attention of generations of researchers and creative writers, including W. H. Auden, Ezra Pound, Seamus Heaney and J. R. R. Tolkien<sup>2,3</sup>.

Within Old English literature, the extant corpus of poetry is relatively small; it comprises around 350 texts, of which over 300 are shorter than 1,000 words in length (Supplementary Fig. 1). The poems are preserved in manuscript copies that provide no direct information about the context in which they originated. Moreover, damage to these manuscripts has frequently resulted in the loss of text, and rendered many poems more or less incomplete. The sparseness and fragmentation of the corpus poses a serious challenge for literary study of the tradition. For instance, it is almost impossible to know how representative, original or popular a particular literary feature might have been when we possess so few comparanda, the authorship or date of extant works is often unknown or uncertain, and the compositional technique of Old English poets is

similarly mysterious. Additional difficulties arise due to uncertainties of register, genre and dialect, which complicate efforts to relate literary works to particular chronological or geographical contexts<sup>4</sup>. Lacking the extensive corpora and contextual evidence that are taken for granted in the study of modern literatures, Old English scholars face considerable difficulties when dealing with questions of literary history. Some scholars have even suggested that the surviving materials are insufficient for meaningful conclusions to be drawn on the basis of linguistic analysis<sup>5</sup>.

One approach to these problems is to extract more information from the material we already have; rather than examining larger and therefore less frequent components of the literature, such as characters or scenes, we can focus on much smaller units, ranging from individual phrases to word segments and even pauses. A benefit of analysing smaller features is that they are necessarily numerous even within a sparse corpus. By combining attention to multiple features of this kind, it is possible to create a high-dimensional profile of a text, or part of a text, in relation to all others in the corpus. Although manual counting of individual small features may be feasible, the generation of high-dimensional profiles generally requires the application of computational techniques. Such techniques have not been employed extensively in the study of Old English literature compared with modern English or even other pre-modern traditions such as Latin<sup>6–9</sup>. Where computation has been brought to bear on Old English texts, the research has generally been limited to a small set of literary features or a handful of specific works<sup>10–15</sup>. The most significant application of modern stylometric techniques to Old English verse has been the development of ‘lexomics’ by Drout et al., which involves the use of vocabulary frequency data and hierarchical clustering to discern literary similarities<sup>12,15</sup>. Lexomic methods have been applied to several important problems, including profiling stylistic differences across *Beowulf* and works associated with Cynewulf (the first author to whom multiple English poems can be attributed)<sup>12,15</sup>. Our methodology complements and extends this earlier work in three principal ways: (1) the use of non-lexical features, especially sense-pauses and metre; (2) attention to specialized word usage, in particular rare nominal compounds; and (3) adaptation of clustering techniques to focus on sequences of characters rather than whole words.

Here, we report a large-scale computational analysis of the entire Old English verse corpus. Our central innovation is to extract information on a wide range of fine-grained features—covering aspects of sound, metre and diction—to discern meaningful stylometric patterns within the corpus as they relate to questions of authorship

<sup>1</sup>Department of English, Nanjing University, Nanjing, China. <sup>2</sup>Program for Evolutionary Dynamics, Harvard University, Cambridge, MA, USA.

<sup>3</sup>Research Science Institute, Center for Excellence in Education, McLean, VA, USA. <sup>4</sup>Texas Academy of Mathematics and Science, Denton, TX, USA.

<sup>5</sup>Department of Classics, University of Texas at Austin, Austin, TX, USA. <sup>6</sup>Department of Systems Biology, Harvard Medical School, Boston, MA, USA.

<sup>7</sup>These authors contributed equally: Leonard Neidorf, Madison S. Krieger. \*e-mail: [mkrieger@fas.harvard.edu](mailto:mkrieger@fas.harvard.edu); [jdexter@fas.harvard.edu](mailto:jdexter@fas.harvard.edu)

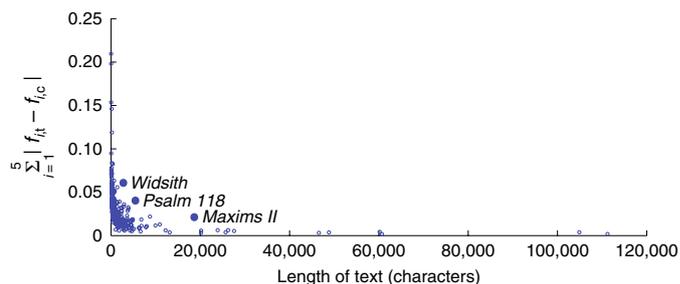
and literary resemblance. Additionally, we introduce a variety of computational tools tailored to the specifics of Old English, which in turn may be valuable in the analysis of other literary and linguistic traditions. We use our corpus-wide profiling to address two longstanding questions in the study of Old English literature: whether *Beowulf* is a unified work of a single author or a combination of multiple texts<sup>16,17</sup>, and whether the anonymous work *Andreas* was written by the poet Cynewulf<sup>18–20</sup>. We show that several orthogonal stylistic metrics do not differ between possible partitions of *Beowulf*, which is consistent with the hypothesis that portions of the poem were not produced separately, or, if they were, that the styles are remarkably uniform. Although this uniformity cannot adjudicate definitively between single or multiple authorship, it militates against a view of the work either as constructed from chronologically disparate poems or as markedly shaped by scribal intervention. Our results also show strong similarity between *Andreas* and other works signed by Cynewulf. Our approach has implications not only for the practice of literary criticism but also for the study of cultural evolution, by generating data on properties of language that probably evolved from this early tradition through Middle and Modern English<sup>21</sup>. While computational analyses cannot definitively resolve longstanding problems arising from a dearth of empirical evidence and theoretical disagreements about cultural production, nevertheless, they do offer additional, quantifiable data that affect the plausibility of various critical hypotheses.

Functional  $n$ -grams are short (typically syllable-length) substrings of natural language text (for example, the substring ‘ab’ in the sentence ‘Abel elaborated about his intentions.’), which have proven useful in previous analyses of both English and Latin literary style, and for authorship attribution, as works by the same author tend to have similar phonetic profiles<sup>9,22–25</sup>. In verse corpora, patterns of functional  $n$ -gram usage can reflect poetic sound play and aural effects. To identify phonetically distinctive poems within the Old English corpus, we computed for each text:

$$\sum_{i=1}^5 |f_{i,t} - f_{i,c}|$$

where  $f_{i,t}$  denotes the frequency of the  $i$ th most common  $n$ -gram in the text, and  $f_{i,c}$  denotes the corpus-wide frequency of that  $n$ -gram. Figure 1 shows a plot of this metric against text length for functional trigrams. Unsurprisingly, numerous short poems appear to have patterns of functional  $n$ -gram usage that differ from the bulk corpus. However, of greater interest is that 3 longer texts (each longer than 125 verses) exhibit unusual patterns of functional trigram usage relative to other texts of comparable length. Profiling of functional bigrams and four-grams similarly identified these same three texts—*Widsith*, *Psalm 118* and *Maxims II*—as anomalous (Supplementary Fig. 2). In other words, these three texts exhibit pronounced deviations from phonetic norms that are otherwise relatively homogeneous throughout the corpus of Old English poetry.

These results prompted us to consider why the phonetics of those three texts should appear distinct from the rest of the corpus. In the case of *Widsith*, the anomaly is likely to be attributed to its preponderance of proper names, which are clustered in 3 lengthy catalogues (lines 18–35, 57–87 and 112–124) that might well have circulated orally before the poem’s composition<sup>26</sup>. If proper names are the cause, the anomalous phonetics of *Widsith* might be an epiphenomenal reflection of a broader phonetic division between the lexicon and the onomasticon of Old English. *Maxims II* shares with *Widsith* the strong possibility that its author drew on pre-existing material, consisting as it does of gnomic statements that could have circulated in smaller or larger catalogues outside the poem. As such, it is plausible that its anomalous  $n$ -gram profile reflects phonetic differences between the archaic constituent material of *Maxims II*

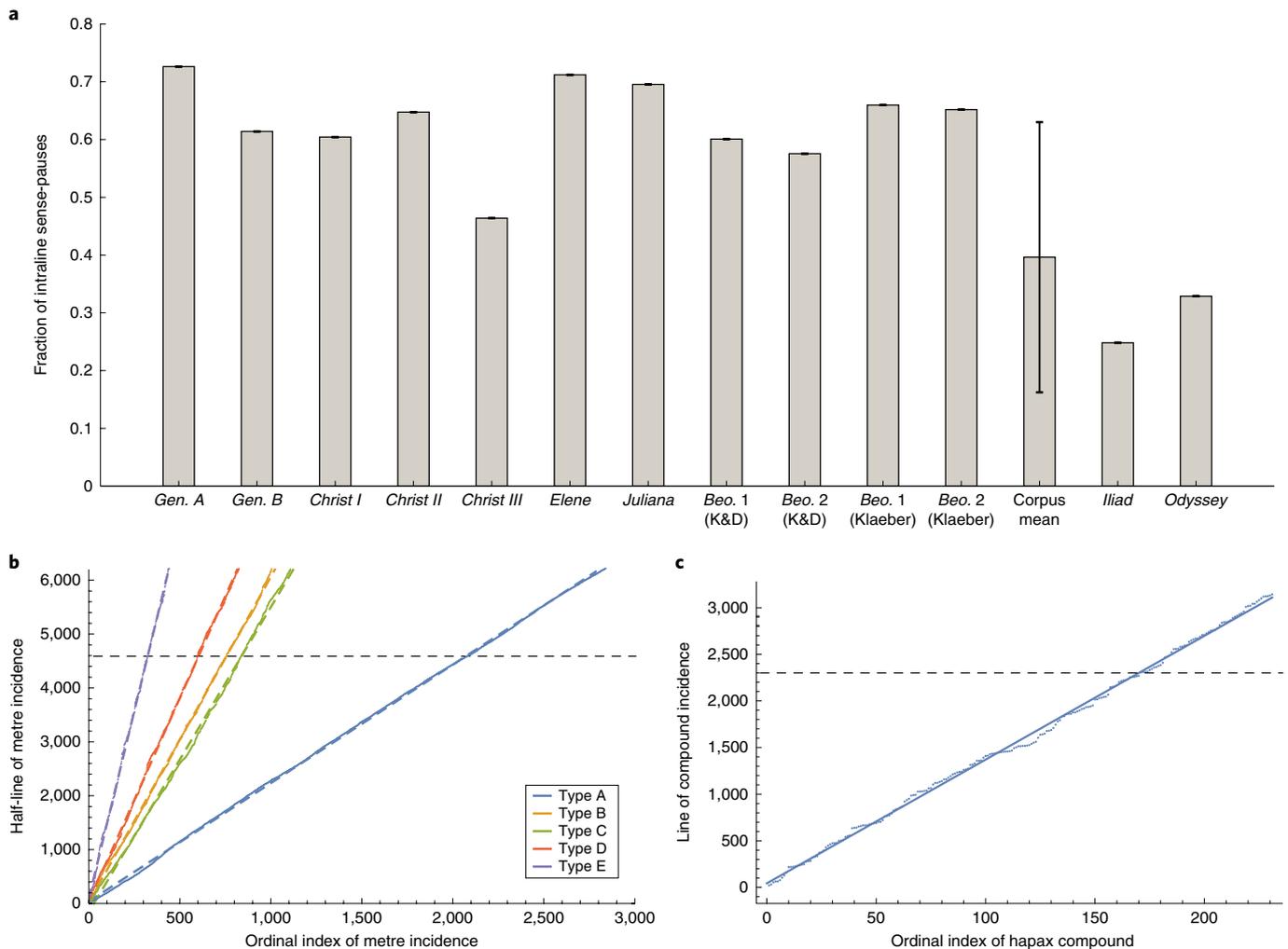


**Fig. 1 | Corpus-wide phonetic profiling of literature.** Plot of the cumulative difference in functional trigram frequency (for the five most common trigrams) against text length. Each dot denotes one text. Three anomalous texts are highlighted and labelled.

and the later linguistic material of which the bulk of the Old English corpus is comprised. *Psalm 118* is peculiar less for its content than for its aberrant metrics and late prosaic vocabulary. There is frequent lexical and syntactic repetition in *Psalm 118*, whereas *Widsith* and *Maxims II* exhibit a far greater degree of structural repetition. As a close translation of a Latin source that might have originated as an interlinear gloss, *Psalm 118* shares with *Widsith* and *Maxims II* the more essential characteristic that its poet’s linguistic freedom was exceptionally constrained by his literary project. Our tests suggest that, under normal conditions, Old English poets generated works that were homogeneous in terms of their phonetic profile. However, this homogeneity was disturbed when a particular literary agenda strongly influenced a poet’s diction or source use.

In addition to analysing the phonetics of the Old English corpus in its entirety, we also sought to address longstanding questions regarding particular texts, beginning with *Beowulf*. Scholarship on *Beowulf* has long entertained debate as to whether the poem is a product of unitary or composite authorship. During the nineteenth century, many prominent scholars subscribed to a theory of composite authorship, which held that *Beowulf* consisted of various pagan lays joined together by Christian editors and interpolators<sup>16</sup>. By the middle of the twentieth century, this view possessed few adherents on account of demonstrations by Klaeber<sup>27</sup> and Tolkien<sup>2</sup> that a coherent Christian perspective pervades the entire poem. Literary critics working in the immediate aftermath of these studies thus tended to premise their work on the assumption that *Beowulf* is the masterwork of a single poet<sup>28,29</sup>. However, theories of composite authorship continued to be propounded throughout the twentieth century, with several scholars arguing that *Beowulf* was put together by a scribal editor who combined two distinct texts: one containing the hero’s fights with Grendel and his mother, and the other containing the hero’s fight with the dragon<sup>30–32</sup>. It has also been argued that scribal interference in the textual transmission of *Beowulf* might have been sufficiently pervasive to render it an essentially composite work<sup>33</sup>. Yet, in the most recent and comprehensive study on the dating and authorship of *Beowulf*, Neidorf<sup>17</sup> adduced a wide range of lexical, metrical, stylistic and palaeographical evidence in support of the contention that the extant manuscript of *Beowulf* faithfully preserves the unitary creation of one poet who composed around the year 700. Here, we offer multiple, orthogonal pieces of quantitative evidence consistent with Neidorf’s view.

As noted above, quantitative analysis of stylistic homogeneity in *Beowulf* has tended to focus on word-based features. To investigate the question further, we devised a broad-spectrum feature set that reflects versification, metre and an aspect of diction (nominal compounds) of particular importance in Old English verse. We first considered sense-pauses, which are breaks in speech typically denoted by any punctuation mark other than a comma. Although sense-pause analysis has not been undertaken previously for Old



**Fig. 2 | Stylistic homogeneity of *Beowulf*.** **a**, Ratio of intraline to total sense-pauses for the partition of *Beowulf*, the mean of all texts in the Old English verse corpus, some salient individual texts, and Homer's *Iliad* and *Odyssey*. The error bar for the corpus mean denotes one s.d. of the ratios for all of the texts. *Beo.*, *Beowulf*; *Gen.*, *Genesis*; K&D, Krapp and Dobbie. **b**, Rate of use of different metres (types A, B, C, D and E) in *Beowulf*. The coloured diagonal lines are linear fits. The dotted horizontal line indicates line 2,300 of the text. **c**, Rate of use of hapax compounds in *Beowulf*. The blue line is a linear fit. The dotted horizontal line indicates line 2,300 of the text.

English literature, it has been applied to questions of stylistic evolution in other traditions. For instance, Fitch<sup>34</sup> demonstrated that the ratio of intraline to total sense-pauses is a reliable marker of relative chronology for the tragedies of Sophocles, Seneca and Shakespeare, perhaps because frequent inclusion of sense-pauses not coincident with line breaks reflects a more confident and mature poetic style.

Theories of composite authorship differ as to the exact division between the component poems, but most suggestions (for example, refs. 16,17) cluster around line 2,300, which is not long after the scribal hand changes in the manuscript (in the middle of line 1,939)<sup>32</sup>. As such, we computed the ratio of intraline to total sense-pauses for lines 1–2,300 and 2,301–end of Krapp and Dobbie's edition of *Beowulf*, along with the corpus-wide average. We calculated that the ratios for lines 1–2,300 and 2,301–end are within 4% of each other (Fig. 2a). As is typical for pre-modern texts, there is no punctuation in the extant Old English manuscripts. However, the editorial judgements about where punctuation is required are guided by various metrical and syntactic regularities, such as those codified in Kuhn's laws, which reliably indicate where clauses begin and end in Old English poetry<sup>35,36</sup>. To account for the remaining freedom in editorial practice regarding punctuation, we analysed another

text of *Beowulf* (edited by Klaeber and revised by Bjork, Fulk and Niles). We found that although the absolute value of the ratios varies between the two editions, the relative difference between lines 1–2,300 and 2,301–end is small in both cases. This comparison suggests that while editorial policies may differ, their consistent application ultimately does not obscure the stylistic regularities in a given poem. The consistency in the handling of intraline sense-pauses across both sections of *Beowulf*, in both editions, therefore provides support for the stylistic unity of the poem.

We sought to corroborate the results of the sense-pause analysis of *Beowulf* through comparison with other Old English poems and with ancient Greek epic. *Genesis*—one of the longest extant Old English poems—is known to be the work of multiple authors; it consists of a later poem, called *Genesis B*, which is approximately 600 lines long and is embedded within the remaining 2,300 or so lines of the older main poem, *Genesis A*. Differences between the two poems have previously been identified using other conventional and quantitative techniques<sup>12</sup>. In our research, we found a marked difference in the intraline-to-total sense-pause ratio between *Genesis A* and *B* (Fig. 2a), suggesting that sense-pause analysis can distinguish between passages of Old English verse about

similar subject matter but composed by different poets. Likewise, the ratio differs between all three *Christ* poems (*Christ I–III*), which are widely held to have been composed by multiple authors. In contrast, we find that the ratio is consistent between *Elene* and *Juliana*, both of which are signed Cynewulfian poems. In aggregate, sense-pause differences are significantly higher in the *Genesis* and *Christ* group than in the *Cynewulf* and *Beowulf* group (two-tailed *t*-test,  $t(5)=2.94$ ;  $P=0.0322$ ; Cohen's  $d=2.25$ ; 95% confidence interval (CI) = 0.013 to 0.194).

Like *Beowulf*, the Greek epics *Iliad* and *Odyssey* have also generated much debate about their authorship and composition. Conventionally attributed to a single author—Homer—both works nevertheless clearly originate in a long oral tradition and show signs of considerable evolution in the course of their transmission history, including the possible influence of written versions<sup>37,38</sup>. Since the two Homeric epics have numerous features in common, we hypothesized that they might also have a similar pattern of sense-pauses. However, as shown in Fig. 2a, the *Odyssey* has a higher proportion of intraline sense-pauses relative to the *Iliad*. This difference suggests a slight change of compositional practice between the two Greek poems, whether due to a single poet's stylistic evolution or natural variation across the oral tradition. Had the two parts of *Beowulf* shown a similar or greater disparity in the sense-pause data when compared with the *Iliad* and the *Odyssey*, this might have supported the view that two different poems had been conjoined. However, as it stands, the comparative uniformity of the data suggests that the compositional practice of both parts was the same, at least with respect to sense-pauses.

We then examined the metre of *Beowulf*. We used a scansion devised by Sievers<sup>39</sup>, which categorizes half-lines into five major sound patterns denoted as types A, B, C, D and E. We investigated both the total frequency of the five verse-types and their sequence within *Beowulf*. Strikingly, we found that the usage rate of each type of metre remains linear across the entirety of *Beowulf* (Pearson's  $r(2,860)=0.998$ ;  $P<0.001$  for type A;  $r(1,008)=0.997$ ;  $P<0.001$  for type B;  $r(1,241)=0.998$ ;  $P<0.001$  for type C;  $r(826)=0.997$ ;  $P<0.001$  for type D;  $r(445)=0.995$ ;  $P<0.001$  for type E), with no discernible shift near line 2,300 and no differences in the frequencies of any particular metrical type (Fig. 2b). To quantify this effect, we computed the difference in slope between the two sections of 1,000 randomly chosen partitions of *Beowulf*; for 4 of the 5 metre types, the mean difference across the partitions is greater than the difference between lines 1–2,300 and 2,301–end (mean = 0.148; s.d. = 0.063 versus 0.0498 for type A; mean = 0.715; s.d. = 0.884 versus 0.331 for type B; mean = 0.717; s.d. = 0.389 versus 0.492 for type C; mean = 0.524; s.d. = 0.697 versus 0.750 for type D; mean = 1.56; s.d. = 1.75 versus 0.575 for type E).

Finally, we considered the distribution of nominal compounds in *Beowulf* and across the Old English verse corpus. Nominal compounds, which are words formed by combining two nouns, are a particularly important aspect of Old English poetry<sup>40</sup>. Examples in Old English include *hron-rad* (whale-road), referring to the sea, and *ban-hus* (bone-house) for the human body. It is generally believed that the number and inventiveness of compounds in a poem is an important marker of literary creativity in the Old English tradition<sup>1,40</sup>. To generate a list of compound words, we identified all entries in the Bosworth–Toller dictionary that are connected with a hyphen and that consist of two separate headwords (hyphens do not appear in native Old English texts)<sup>41</sup>. Supplementary Fig. 3 shows the distribution of compound words by frequency of occurrence. For our initial analysis, we considered inter-authorial differences in the usage of hapax legomena compound words (that is, compounds that appear only once in the entire poetic corpus). The rate of usage of hapax compounds can be very different between authors, as illustrated in Supplementary Fig. 4 for two of the longest extant poems (*Genesis* and *Exodus*). As discussed above, *Genesis* is known to be a

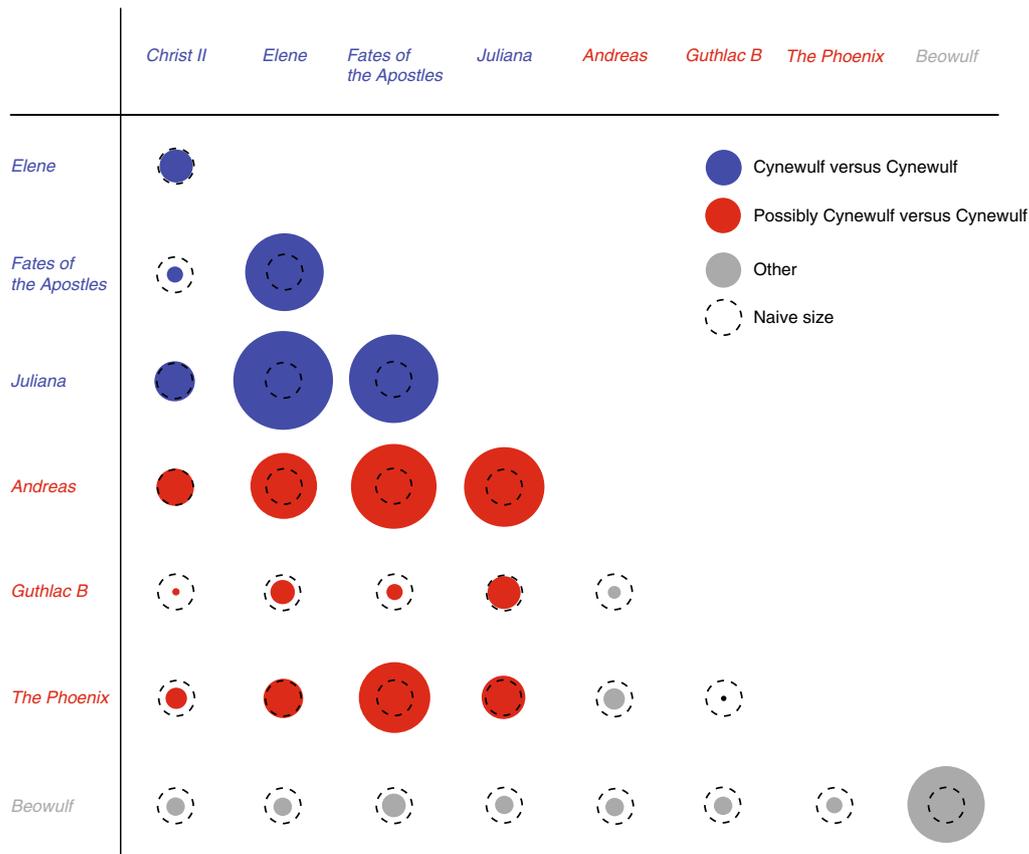
composite work. We partitioned *Genesis A* into two random sections that are of comparable length to *Genesis B* and analysed the rate of hapax usage. Linear fits to the data for both sections of *Genesis A* have very similar slopes and differ from the fit to the *Genesis B* data (Supplementary Fig. 4a). In contrast, *Exodus*—the unitary authorship of which has never been in dispute—shows clear homogeneity when analysed in the same way (Supplementary Fig. 4b).

Taken together, these results demonstrate that nominal compounds are an effective metric for Old English stylistic and attribution studies. We therefore constructed a profile of hapax compounds across the whole of *Beowulf* (Fig. 2c). This profile revealed that the rate of compound usage is linear throughout the poem (Pearson's  $r(229)=0.992$ ;  $P<0.001$ ), with no change in slope observed around line 2,300. The difference in slope between lines 1–2,300 and 2,301–end is 1.50 (mean = 1.42; s.d. = 1.02 for 1,000 random partitions). The small nonlinearity evident around line 1,500 corresponds to Beowulf's fight with Grendel's mother, which is known to be particularly rich in compound words and other distinctive linguistic features. Accordingly, our analysis of nominal compounds provides further evidence (orthogonal to the sense-pause and metrical data) for the stylistic homogeneity of *Beowulf*.

Our other major results concern a collection of poems written by an author called Cynewulf, or by a broader 'Cynewulfian school'. Four Old English poems—*Elene*, *Juliana*, *Christ II* and *Fates of the Apostles*—conclude with epilogues that ascribe their composition to an otherwise unknown individual named Cynewulf. Many scholars, perceiving stylistic or thematic affinities between these signed works and other anonymous poems, have sought to expand Cynewulf's corpus to include such works as *Andreas*, *Guthlac A/B*, *Christ I/III*, *Judith*, *The Phoenix* and *The Dream of the Rood*, among other poems<sup>42–44</sup>. While once considered products of Cynewulf's own hand, these poems are now more commonly regarded as products of a Cynewulfian school of poetry, if they are believed to possess any meaningful connections to his work at all<sup>45</sup>. The majority of scholars in the past half-century consider Cynewulf to be the author of only the four signed poems, although some have maintained that either *Guthlac B*, *Andreas* or both should be included in his corpus as well, in part based on computational stylometric analysis<sup>12,46</sup>. In addition, whether Cynewulf should even be regarded as the author of the four poems bearing his signature has been questioned, since it is theoretically possible that Cynewulf added his epilogues to poems that other authors originally composed<sup>20,47</sup>. Our tests assuage such doubts by identifying a strong degree of stylistic homogeneity among three of the four signed works of Cynewulf. This homogeneity supports the longstanding assumption that one author composed at least three, and possibly all four, of the poems in question. We also find compelling evidence for an association between *Andreas* and Cynewulf's poetry, which might indicate—in contrast with current opinion—that Cynewulf composed this poem as well.

We first compared the usage of hapax compounds across ten Old English poems, including three control texts not by Cynewulf (*Beowulf*, *Exodus* and *Christ and Satan*), the four signed Cynewulf poems and three poems often associated with Cynewulf (*Andreas*, *Guthlac B* and *The Phoenix*) (Supplementary Fig. 4c). The three control poems, which are thought to be by different authors writing during different periods in Anglo-Saxon history, unsurprisingly show distinct patterns of compound usage. However, the signed Cynewulf poems appear similar both to each other (although *Christ II* shows less affiliation with the other works) and to *Andreas*.

This result prompted us to examine the similarity of *Andreas* to the signed poems of Cynewulf on the basis of a broader range of nominal compounds beyond hapax legomena. In Old English, multiple compounds could denote a single object or concept. There are at least 17 completely distinct compound words in the poetic corpus that denote 'the sea', for example, and 11 compounds meaning 'warrior'<sup>48,49</sup>. Therefore, an author's particular choice of compound

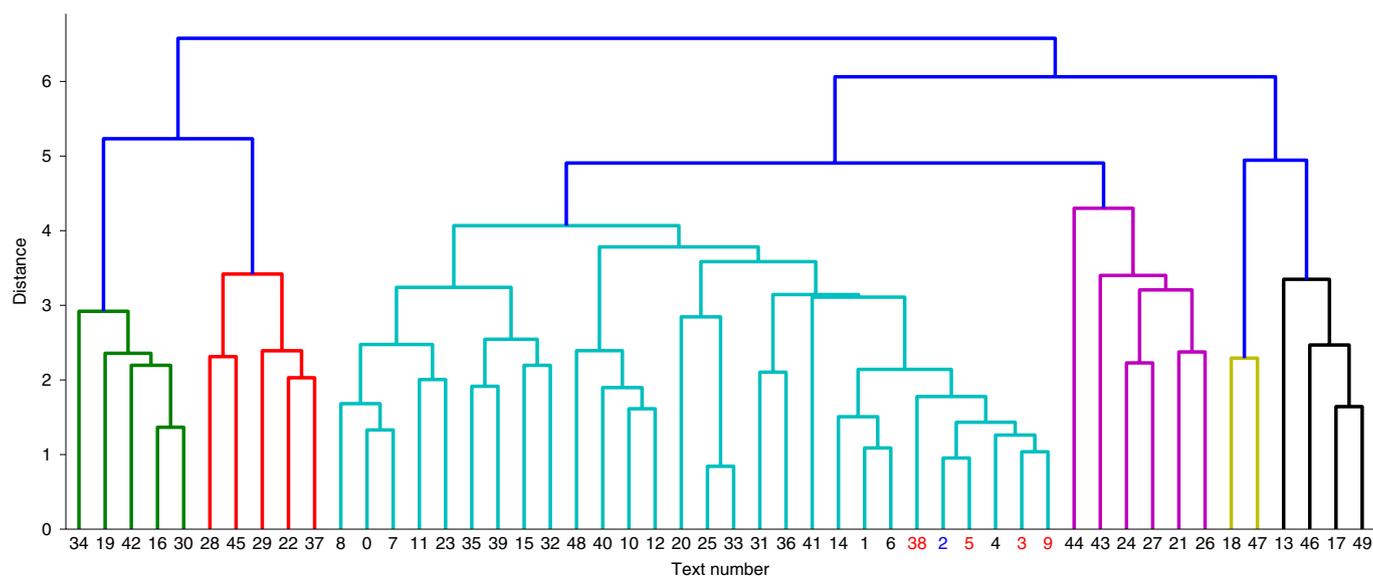


**Fig. 3 | Use of nominal compounds is similar between Cynewulf and Andreas.** Distribution of non-unique compounds in six poems either signed by Cynewulf (blue) or of possible Cynewulfian authorship (red) and in *Beowulf* (grey). The size of each filled circle indicates the number of compounds shared by the corresponding pair of texts. The dotted open circles indicate the expected size if all compounds were distributed at random. The circle in the bottom right is for comparison of *Beowulf* 1–2,300 and *Beowulf* 2,301–end.

might reflect a variety of factors: nuance in meaning, literary influence or other linguistic considerations. Accordingly, the usage of compounds forms an important part of the stylistic profile of an Old English author. We performed a large-scale analysis of non-hapax compounds (excluding only *wuldorcyning*, *heofoncyning* and *heofonrice* ('wonder-king', 'heaven-king' and 'heaven-kingdom', respectively), which occur with extremely high frequency throughout much of the Old English religious poetry) in the verse corpus (Fig. 3). Each solid circle in Fig. 3 denotes the degree of correlation between the two indicated texts, compared with a random distribution of compound words based on their overall frequency and the lengths of the individual poems (dotted circles). By this measure, most of the signed works of Cynewulf are strongly correlated. However, *Christ II* is close to naively correlated, perhaps due to an absence of the compounds that are typically used in the hagiographical poems (*Elene* and *Juliana*) to express the divine relationship between the saint and God (for example, *mundbyrd* ('suffrage/aid')). By the same measure, *Andreas* is strongly correlated with the signed poems of Cynewulf, in agreement with our analysis of hapax compounds (Supplementary Fig. 4). Supporting these observations, at least one of the Cynewulf/Cynewulf (blue circles), Cynewulf/*Andreas* (top line of red circles) and Cynewulf/other (remaining red circles) comparison groups is significantly different from the others (one-way analysis of variance,  $F(2,18) = 5.73$ ;  $P = 0.0119$ ). There is no significant pairwise difference between Cynewulf/Cynewulf and Cynewulf/*Andreas* ( $Q(18) = 0.653$ ;  $P = 0.890$ ; Cohen's  $d = -0.244$ ; 95% CI =  $-0.952$  to  $1.37$ ), but there is a significant difference between Cynewulf/Cynewulf

and Cynewulf/other ( $Q(18) = 3.75$ ;  $P = 0.0410$ ; Cohen's  $d = 1.32$ ; 95% CI =  $-1.86$  to  $-0.0354$ ) and between Cynewulf/*Andreas* and Cynewulf/other ( $Q(18) = 3.98$ ;  $P = 0.0294$ ; Cohen's  $d = 2.04$ ; 95% CI =  $-2.21$  to  $-0.108$ ; all values are from post-hoc Tukey–Kramer tests). Additionally, we observe that *Beowulf* is self-correlated when partitioned into lines 1–2,300 and 2,301–end, which provides further support for unitary composition.

Finally, we used hierarchical agglomerative clustering to investigate the possible association of *Andreas* with the signed poems of Cynewulf on the basis of functional  $n$ -gram frequencies, which are often used for authorship attribution studies involving literary texts written in Modern English<sup>22,24</sup>. As described in detail in the Methods, we computed the frequencies of the 25 most common trigrams (based on the corpus-wide frequency) in the 50 longest poems, with *Beowulf* partitioned into 2 parts as usual. We used hierarchical agglomerative clustering with this feature set to construct the dendrogram shown in Fig. 4. In line with our other studies, *Beowulf* lines 1–2,300 and 2,301–end cluster together. Furthermore, we find that *Andreas* clusters next to *Elene* and in close proximity to *Juliana*, *Fates of the Apostles* and *Christ I/II/III*. Also in this cluster is *Guthlac A* and *B*, the latter of which Drout et al. associated with the works of Cynewulf based on a clustering analysis with word-level features and a small subset of the Old English verse corpus<sup>12</sup>. To investigate the robustness of these observations, we repeated the clustering with bigrams and four-grams and found that key aspects of the dendrogram structure, including the side-by-side positioning of the two parts of *Beowulf*, and the positioning of *Andreas* next to *Elene* and in close proximity to at least two other signed



**Fig. 4 | Andreas clusters with the signed Cynewulfian poems.** Dendrogram generated from hierarchical agglomerative clustering of the 50 longest poems, with high-frequency functional trigrams as features. The text numbers are given in Supplementary Table 1; for reference, *Andreas* is text 2 (blue) and the Cynewulfian poems are all red (*Christ*, 3; *Elene*, 5; *Juliana*, 9; and *Fates of the Apostles*, 38). *Beowulf* 1–2,300 is 1, and *Beowulf* 2,301–end is 6.

Cynewulfian poems, were preserved in both cases (Supplementary Fig. 5). Our results, obtained using unsupervised learning and a type of feature (high-frequency functional  $n$ -grams) well-established in the attribution literature, thus corroborate the stylistic association between Cynewulf and *Andreas* that we identified through analysis of nominal compounds.

The tests we have conducted indicate some ways in which quantitative profiling of the Old English verse tradition can help to answer or raise questions of considerable interest to researchers. With regard to *Beowulf*, our tests tilt the scales of probability between hypotheses that are currently in competition. In contrast, with Cynewulf, our tests encourage scholars to reconsider a possibility that has not been seriously entertained in the past half-century. Our evidence for the stylistic homogeneity of *Beowulf* does not prove that the poem is the work of one individual, but it substantially enhances the probability of unitary authorship, while presenting serious obstacles to those who would advocate for composite authorship or scribal recomposition. Our evidence for the extraordinary affinities between the language of *Andreas* and the language of the signed works of Cynewulf similarly does not prove that *Andreas* was composed by Cynewulf, but it demands that this possibility be explored further in future studies. Orchard, noticing many formulaic expressions shared between *Andreas* and the works of Cynewulf, interpreted the overlap as an indication that the *Andreas* poet read the works of Cynewulf and borrowed extensively from them<sup>19</sup>. Given the lack of decisive evidence, we must acknowledge the possibility, both for *Beowulf* and *Andreas*, that some combination of generic constraints and highly skilled imitation might account for the patterns observed. In each case, however, the most economical explanation of the data is similar: unitary composition of *Beowulf* and Cynewulfian authorship of *Andreas*. Furthermore, in view of the fact that *Andreas* immediately precedes *Fates of the Apostles* in the Vercelli Book (the manuscript in which these two poems are preserved), we might tentatively regard Cynewulf's signature at the end of the latter as a claim to authorship of the former as well.

Our results show the utility of taking a wide range of quantitative approaches to the study of a literary corpus, from simple frequency counts to machine learning. However, crucial to the success of any large-scale profiling is the selection of features used to characterize

the corpus<sup>50</sup>. In this case, the variety of features complements and enhances the more established focus on word usage and distribution, incorporating in addition phonetic, formulaic, rhythmic and metrical elements. In doing so, we exploit features that are known to play an important role in the specific tradition (for example, nominal compounds), as well as validate the extension of features that have proven useful for studying traditions in other languages (for example, functional  $n$ -grams and sense-pauses) to Old English<sup>9,23,34</sup>. In our analysis of Cynewulf, we show that a corpus-specific feature (nominal compounds) can be combined with a general-purpose stylistic technique (unsupervised learning with character  $n$ -grams) to provide broad-based support for the Cynewulfian authorship of *Andreas*. Moreover, the quantitative tests designed to analyse nominal compounds might be profitably applied in the future to other languages and traditions where aspects of word formation allow for the free combination of simpler lexical items into larger, often unique units, such as agglutinative and polysynthetic languages<sup>51</sup>. In summary, our diverse combination of methods and features constitutes an effective response to the challenges posed by sparse corpora. In particular, the computational analysis of many microscopic features yields results that either cannot be obtained using conventional critical methods or can only be obtained with great difficulty.

Our approach provides a model applicable to other literary traditions. Although potentially useful for the analysis of any corpus of literature, the techniques described here offer a particular advantage for the study of corpora posing similar challenges as Old English poetry, such as other medieval traditions including Old Norse, Old Irish and Old French<sup>52–54</sup>. These languages exhibit many characteristics shared with Old English and are hence especially amenable to the same methods. However, all pre-modern literary traditions suffer to a greater or lesser extent from the problem of text loss, and hence sparse corpora—a situation compounded by the frequent lack of contextual information about the date or authorship of works. Our study suggests some general ways of overcoming or circumventing these challenges, and of finding data that can shed light on both work-specific and corpus-wide questions. Generating quantitative profiles for multiple literary traditions would also represent an initial step towards a quantitative analysis of literature

across cultures. Furthermore, in focusing on a pre-modern tradition—especially one that has seen relatively little computational research—our work broadens the digital humanities' predominant concern with modern literature, and lays the foundation for future diachronic profiling of the English literary tradition with substantial time depth<sup>55</sup>.

## Methods

**Corpora and text processing.** The texts of the Old English verse corpus were obtained from the University of Calgary's Online Corpus of Old English Poetry (OCOEP) in UTF-8 encoding (<http://www.oepoetry.ca/>), which preserves native Old English characters and contains character markings separating half-lines and full-lines, as well as different poems. Except in the following two cases, we used the complete, unaltered OCOEP for corpus-wide analyses: (1) before computing the corpus mean for sense-pauses (Fig. 2a), we aggregated related short texts (for example, the poems of the *Paris Psalter* and the *Meters of Boethius* into single files; and (2) we restricted the hierarchical clustering analysis to the 50 longest texts in the unaltered corpus, with the two partitions of *Beowulf* counted separately (Fig. 4 and Supplementary Fig. 5). Greek texts of the *Iliad* and *Odyssey* were obtained from the Tesseract Project, whose corpus is derived from the Perseus Digital Library (<http://www.perseus.tufts.edu/hopper/>).

**Natural language processing.** All natural language processing tasks were performed using Python 3.6.4.

**Calculation of sense-pause frequency.** Following the definition of Fitch for Greek, Latin and modern English poetry<sup>34</sup>, we determined sense-pause frequencies by tabulation of punctuation marks other than commas (., ?, !, ;, :, (, ), -, ' , " and "). Any punctuation mark not coincident with a line break was considered to be an intraline sense-pause.

**Metrical analysis.** To supplement the OCOEP text file of the corpus for metrical analysis, we sought scansion of the longest poems, which were provided by G. Russom<sup>56</sup>. We then identified for *Beowulf* the total frequency of each of the five verse-types, as well as the half-lines on which they occurred.

**Identification of nominal compounds.** We compiled a list of compound words from the set of hyphenated noun–noun headwords in the online Bosworth–Toller dictionary<sup>31</sup>, excluding only one compound (*middangeard*, which means 'middle-land' or 'Earth' and inspired Tolkien's 'Middle Earth'). This compound is used with unusually high frequency (135 instances) and appears to have been used in a manner distinct from other poetic compounds throughout most of the Old English literary period. For each compound in the list, we identified the set of all poems in which that compound occurs, which was used to generate Fig. 2c and Supplementary Fig. 4. We also computed a measure of correlation between poems, defined relative to a random distribution of compound words according to a compound's frequency and the length of the poem. This random distribution of compound words was calculated 10,000 times per compound word to generate a well-defined distribution over the corpus for that word. For Fig. 3, we summed all of the compound words that appear in multiple poems, which quantifies the extent to which each pair of poems has shared compounds. The radius of each circle in Fig. 3 is the ratio of this number to the number predicted by the random distribution.

**Hierarchical agglomerative clustering.** To generate feature sets for clustering analysis, we determined the 25 most common functional bigrams, trigrams and four-grams in the Old English verse corpus and computed their frequency in the 50 longest poems (with *Beowulf* partitioned into lines 1–2,300 and 2,301–end). We used the scipy implementation of hierarchical agglomerative clustering with the Euclidean distance metric and Ward's linkage criterion to cluster those 50 texts. Dendrograms for the bigram and four-gram clustering are shown in Supplementary Fig. 5, and a dendrogram for the trigram clustering is shown in Fig. 4.

**Reporting Summary.** Further information on research design is available in the Nature Research Reporting Summary linked to this article.

## Data availability

All datasets are freely and publicly available at <https://github.com/qcrist>.

## Code availability

All custom code is freely and publicly available at <https://github.com/qcrist>.

Received: 23 November 2017; Accepted: 3 March 2019;

Published online: 8 April 2019

## References

- Fulk, R. & Cain, C. A *History of Old English Literature* 2nd edn (Wiley-Blackwell, 2013).
- Tolkien, J. *Beowulf: the monsters and the critics*. *Proc. Br. Acad.* **22**, 245–295 (1936).
- Clark, D. & Perkins, N. *Anglo-Saxon Culture and the Modern Imagination* (D. S. Brewer, 2010).
- Biber, D. & Conrad, S. *Register, Genre, and Style* (Cambridge Univ. Press, 2009).
- Amos, A. C. *Linguistic Means of Determining the Dates of Old English Literary Texts* (Medieval Academy of America, 1980).
- Jockers, M. *Macroanalysis: Digital Methods and Literary History* (Univ. Illinois Press, 2013).
- Long, H. & So, R. Literary pattern recognition: modernism between close reading and machine learning. *Crit. Inq.* **42**, 235–267 (2016).
- Chaudhuri, P. & Dexter, J. P. Bioinformatics and classical literary study. *Journal of Data Mining & Digital Humanities Special Issue on Computer-Aided Processing of Intertextuality in Ancient Languages* <https://jdm.dh.episciences.org/paper/view?id=3807> (2017).
- Dexter, J. et al. Quantitative criticism of literary relationships. *Proc. Natl Acad. Sci. USA* **114**, E3195–E3204 (2017).
- Barquist, C. Phonological patterning in *Beowulf*. *Lit. Linguist. Comput.* **2**, 19–23 (1987).
- Barquist, C. & Shie, D. Computer analysis of alliteration in *Beowulf* using distinctive feature theory. *Lit. Linguist. Comput.* **6**, 274–280 (1991).
- Drout, M. D., Kahn, M. J., LeBlanc, M. D. & Nelson, C. Of dendrogrammar: lexicomic methods for analyzing relationships among Old English poems. *J. Eng. Ger. Philol.* **110**, 301–336 (2007).
- García, A. M. & Martín, J. C. Function words in authorship attribution studies. *Lit. Linguist. Comput.* **22**, 49–66 (2007).
- Gill, P., Swartz, T. & Treschow, M. A stylometric analysis of King Alfred's literary works. *J. Appl. Stat.* **34**, 1251–1258 (2007).
- Drout, M., Kisor, Y., Smith, L., Dennett, A. & Piirainen, N. *Beowulf Unlocked: New Evidence from Lexomic Analysis* (Palgrave Macmillan, 2016).
- Shippey, T. in *A Beowulf Handbook* (eds Bjork, R. & Niles, J.) 159–168 (Univ. Nebraska Press, 1998).
- Neidorf, L. *The Transmission of Beowulf* (Cornell Univ. Press, 2017).
- Bjork, R. *Cynewulf: Basic Readings* (Garland Publishing, 1996).
- Orchard, A. in *Anglo-Saxon Styles* (eds Karkov, C. & Brown, G.) 271–305 (State Univ. New York Press, 2003).
- Puskar, J. Questioning Cynewulf's claim of authorship. *Eng. Stud.* **92**, 1–19 (2011).
- Mesoudi, A. Pursuing Darwin's curious parallel: prospects for a science of cultural evolution. *Proc. Natl Acad. Sci. USA* **114**, 7853–7860 (2017).
- Grieve, J. Quantitative authorship attribution: an evaluation of techniques. *Lit. Linguist. Comput.* **22**, 251–269 (2007).
- Forstall, C., Jacobson, S. & Scheirer, W. Evidence of intertextuality: investigating Paul the Deacon's *Angustae Vitae*. *Lit. Linguist. Comput.* **26**, 285–296 (2011).
- Koppel, M., Schler, J. & Argamon, S. Computational methods in authorship attribution. *J. Assoc. Inf. Sci. Technol.* **60**, 9–26 (2009).
- Sapkota, U., Bethard, S., Montes, M. & Solorio, T. Not all character *n*-grams are created equal: a study in authorship attribution. In *Proc. 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* 93–102 (Association for Computational Linguistics, 2015).
- Neidorf, L. The dating of *Widsith* and the study of Germanic antiquity. *Neophilologus* **97**, 165–183 (2013).
- Klaeber, F. *The Christian Elements in Beowulf* (Medieval Institute Publications, 1996).
- Brodeur, A. *The Art of Beowulf* (Univ. California Press, 1959).
- Irving, E. *A Reading of Beowulf* (Yale Univ. Press, 1968).
- Schücking, L. L. *Beowulfs Rückkehr: Eine Kritische Studie* (M. Niemeyer, Halle, 1906).
- Magoun, F. P. in *Early English and Norse Studies Presented to Hugh Smith in Honour of his Sixtieth Birthday* (eds Brown, A. & Foote, P.) 127–140 (Methuen, 1963).
- Kiernan, K. S. *Beowulf and the Beowulf Manuscript* (Rutgers Univ. Press, 1981).
- Liuzza, R. M. in *Beowulf: Basic Readings* (ed. Baker, P.) 281–302 (Garland Publishing, 1995).
- Fitch, J. Sense-pauses and relative dating in Seneca, Sophocles and Shakespeare. *Am. J. Philol.* **102**, 289–307 (1981).
- Kuhn, H. Zur Wortstellung und -betonung im Altgermanischen. *Beitr. Gesch. Dtsch. Sprache Lit.* **57**, 1–109 (1933).
- Momma, H. The composition of Old English poetry. *Lang. Lit.* **7**, 175–178 (1998).
- Nagy, G. *Homer's Text and Language* (Univ. Illinois Press, 2004).
- West, M. L. *The Making of the Iliad: Disquisition and Analytical Commentary* (Oxford Univ. Press, 2011).

39. Sievers, E. *Altgermanische Metrik* (M. Niemeyer, Halle, 1893).
40. Gardner, T. The Old English kenning: a characteristic feature of Germanic poetical diction? *Mod. Philol.* **67**, 109–117 (1969).
41. Bosworth, J. *An Anglo-Saxon Dictionary* (Clarendon Press, 1989).
42. Cook, A. S. *The Christ of Cynewulf: A Poem in Three Parts: the Advent, the Ascension and the Last Judgment* (Ginn and Company, 1900).
43. Diamond, R. E. The diction of signed poems in Cynewulf. *Philolog. Q.* **38**, 228–241 (1959).
44. Schaar, C. *Critical Studies in the Cynewulf Group* (Haskell House, 1967).
45. Fulk, R. in *Cynewulf: Basic Readings* (ed. Bjork, R. E.) 3–22 (Garland Publishing, 1996).
46. Bjork, R. E. *The Old English Poems of Cynewulf* (Cambridge Univ. Press, 2013).
47. Stodnick, J. A. Cynewulf as author: medieval reality or modern myth? *Bull. J. Rylands Univ. Libr.* **79**, 25–39 (1997).
48. Carr, C. T. *Nominal Compounds in Germanic* (St. Andrews Univ., 1939).
49. Terasawa, J. *Nominal Compounds in Old English: A Metrical Approach* (Rosenkilde & Bagger, 1994).
50. Jockers, M. L. & Underwood, T. in *A New Companion to Digital Humanities* 2nd edn (eds Schreibman, S., Siemens, R. & Unsworth, J.) 291–306 (Wiley-Blackwell, 2016).
51. Greenberg, J. H. A quantitative approach to the morphological typology of language. *Int. J. Am. Ling.* **26**, 178–194 (1960).
52. O'Donoghue, H. *Old Norse-Icelandic Literature: A Short Introduction* (Wiley-Blackwell, 2004).
53. Bhrolchain, M. N. *An Introduction to Early Irish Literature* (Four Courts Press, 2017).
54. Zink, M. *Medieval French Literature: An Introduction* (Medieval and Renaissance Texts and Studies, 1995).
55. Dimock, W. C. Low epic. *Crit. Inq.* **39**, 614–631 (2013).
56. Russom, G. *Old English Meter and Linguistic Theory* (Cambridge Univ. Press, 1987).

## Acknowledgements

The authors thank M. Nowak, S. Sinai and J. Gerold for helpful conversations, as well as S. Pintzuk and G. Russom for assistance in obtaining texts, dictionaries and scansions in formats amenable to computational analysis. This work was conducted under the auspices of the Quantitative Criticism Lab ([www.qcrit.org](http://www.qcrit.org)), an interdisciplinary project co-directed by P.C. and J.P.D. and supported by a Neukom Institute for Computational Science CompX Grant and a National Endowment for the Humanities Digital Humanities Start-Up Grant (HD-248410-16). P.C. was supported by a New Directions Fellowship from the Andrew W. Mellon Foundation, and J.P.D. was supported by a National Science Foundation Graduate Research Fellowship (DGE1144152) and a Neukom Fellowship. The Program for Evolutionary Dynamics is supported in part by a gift from B. Wu and E. Larson. The funders had no role in study design, data collection and analysis, decision to publish or preparation of the manuscript.

## Author contributions

L.N., M.S.K., P.C. and J.P.D. designed the study. M.S.K., M.Y. and J.P.D. performed the study. All authors analysed the results. L.N., M.S.K., P.C. and J.P.D. wrote the manuscript, which was read and approved by all authors.

## Competing interests

The authors declare no competing interests.

## Additional information

**Supplementary information** is available for this paper at <https://doi.org/10.1038/s41562-019-0570-1>.

**Reprints and permissions information** is available at [www.nature.com/reprints](http://www.nature.com/reprints).

**Correspondence and requests for materials** should be addressed to M.S.K. or J.P.D.

**Publisher's note:** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

© The Author(s), under exclusive licence to Springer Nature Limited 2019

## Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see [Authors & Referees](#) and the [Editorial Policy Checklist](#).

### Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a Confirmed

- The exact sample size ( $n$ ) for each experimental group/condition, given as a discrete number and unit of measurement
- A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
- The statistical test(s) used AND whether they are one- or two-sided  
*Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
- A description of all covariates tested
- A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
- A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
- For null hypothesis testing, the test statistic (e.g.  $F$ ,  $t$ ,  $r$ ) with confidence intervals, effect sizes, degrees of freedom and  $P$  value noted  
*Give  $P$  values as exact values whenever suitable.*
- For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
- For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
- Estimates of effect sizes (e.g. Cohen's  $d$ , Pearson's  $r$ ), indicating how they were calculated

*Our web collection on [statistics for biologists](#) contains articles on many of the points above.*

### Software and code

Policy information about [availability of computer code](#)

Data collection

Data analysis

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research [guidelines for submitting code & software](#) for further information.

### Data

Policy information about [availability of data](#)

All manuscripts must include a [data availability statement](#). This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

## Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

- Life sciences       Behavioural & social sciences       Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see [nature.com/documents/nr-reporting-summary-flat.pdf](https://www.nature.com/documents/nr-reporting-summary-flat.pdf)

## Behavioural & social sciences study design

All studies must disclose on these points even when the disclosure is negative.

Study description	The study was mixed-methods, involving quantitative text analysis followed by qualitative literary critical and philological interpretation.
Research sample	The research sample was texts contained in the University of Calgary's Online Corpus of Old English Poetry, which is freely and publicly available at <a href="http://www.oepoetry.ca/">http://www.oepoetry.ca/</a> .
Sampling strategy	No sampling was done; all results are based on exact calculations using some or all of the Online Corpus of Old English Poetry.
Data collection	N/A
Timing	N/A
Data exclusions	N/A
Non-participation	N/A
Randomization	N/A

## Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

### Materials & experimental systems

n/a	Involvement in the study
<input checked="" type="checkbox"/>	<input type="checkbox"/> Antibodies
<input checked="" type="checkbox"/>	<input type="checkbox"/> Eukaryotic cell lines
<input checked="" type="checkbox"/>	<input type="checkbox"/> Palaeontology
<input checked="" type="checkbox"/>	<input type="checkbox"/> Animals and other organisms
<input checked="" type="checkbox"/>	<input type="checkbox"/> Human research participants
<input checked="" type="checkbox"/>	<input type="checkbox"/> Clinical data

### Methods

n/a	Involvement in the study
<input checked="" type="checkbox"/>	<input type="checkbox"/> ChIP-seq
<input checked="" type="checkbox"/>	<input type="checkbox"/> Flow cytometry
<input checked="" type="checkbox"/>	<input type="checkbox"/> MRI-based neuroimaging