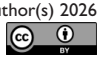
Journal of English Linguistics
2026, Vol. 54(2) 83–108



DOI: 10.1177/00754242261426400
journals.sagepub.com/home/eng


Issues of Faithfulness and Sampling in Historical Corpora: Challenges and Recommendations

James M. Stratton[1]

Abstract

Electronically available corpora have inevitably enhanced access to historical texts, but building historical corpora can often, unintentionally, distort historical language. This paper discusses how issues of faithfulness and sampling can impact the reliability of a historical corpus-based analysis, with recommendations provided for both corpus users and corpus builders. First, by comparing the Beowulf manuscript with the language found in a historical corpus, this paper shows how seemingly innocent editorial changes such as spacing, punctuation, and spelling can have linguistic consequences for phonological, morphological, syntactic, and socio-historical analyses. The removal or addition of elements is argued to have both theoretical and methodological implications. Second, this paper discusses how sampling issues can lead to erroneous conclusions by corpus users about historical language. Recommendations are provided regarding supplementation and mixed-method approaches.

Keywords

historical corpora, faithfulness, sampling, Beowulf, manuscripts, Old English *fela*

1. Introduction

The availability of historical corpora has revolutionized the way linguistic analyses can be conducted, with the once strenuous task of traveling to archives and laboriously gaining permission to sift through century-old manuscripts, being replaced, in theory, with a corpus search query. With machine-readable versions of historical texts at their

[1]Penn State, University Park, PA, USA

Corresponding Author:
James M. Stratton, Penn State, University Park, PA 16802-1503, USA.
Email: james.stratton@psu.edu

fingertips, researchers have a wealth of language material that decades ago would have been unmatched. However, despite the plentiful benefits that historical corpora can provide, various pitfalls can ensue when original manuscripts, or high-resolution facsimiles thereof, are not consulted, leading to erroneous and sometimes anachronistic assumptions and conclusions about socio-historical variation and change. Sampling issues can also distort historical reality, painting a skewed picture of variation, with some varieties overrepresented and others underrepresented, marginalized, or simply not considered.

This article elaborates on two specific challenges that can arise when using and building historical corpora: faithfulness and sampling, with a particular emphasis on the early history of English. In addition to discussing how these two factors can influence our contemporary understanding of historical language and society, recommendations are provided for both corpus users and corpus compilers. As texts are edited and corpus compilers decide what or what not to include in a corpus, information can be lost, misinterpreted, altered, or dismissed. While some editorial decisions may be designed to ease the readability of a text for modern readers, textual alterations can shape users' views about historical variation and change.

The structure of this paper is as follows. Section 2 discusses how editorial intervention can lead to erroneous assumptions, interpretations, and conclusions about language variation and change, with seemingly minor editorial changes such as spacing, punctuation, and spelling having the potential for theoretical and methodological implications. Section 3 addresses sampling issues in historical corpora, with discussions on when and how to supplement corpus data. Both sections end with recommendations for corpus users and compilers. A summary of the major points is provided in the conclusion in Section 4.

2. Check the Original Manuscript and Your Assumptions

2.1. Overview

In 1989, Matti Rissanen called attention to the "philologist's dilemma," one of three issues he foresaw in the use of diachronic corpora. The philologist's dilemma refers to the "risk that corpus work and computer-supported quantitative research methods will discourage the student from getting acquainted with original texts" (Rissanen 1989:16). The solution, he claimed, was to constantly remind ourselves to read the originals (Rissanen 1989:16). But an obvious question is, why? Against the backdrop of the philologist's dilemma and to provide one possible answer, this section compares, as a case study, the language of Beowulf, as attested in the original manuscript (Cotton Vitellius A. XV. MS), with "edited editions"[1] of the text: Beowulf in the *Helsinki Corpus of English Texts* (Kytö 1991; Rissanen, Kytö & Palander-Collin 1993) and two textual editions of the poem (Alexander 2005; Fulk et al. 2008), drawing attention to issues of faithfulness in edited and corpus-based editions.

Faithfulness, that is, how much a text in a corpus resembles its original form, is a constraint that is well documented in corpus linguistics (Lass 2004; Curzan & Palmer 2006; Grund 2006; Horobin 2012; Kytö & Pahta 2012; Stratton 2020a). Although it is acknowledged that most historical texts in a corpus have usually undergone some degree of editorial intervention (Grund, Kytö & Rissanen 2004; Horobin 2012:53-57; Kytö & Pahta 2012:125), it is unclear how cognizant users are of the extent to which texts in a historical corpus may differ from the originals. By comparing the extant copy of Beowulf with the edited and corpus-based editions, this section hopes to reinforce the call for returning to the original manuscripts when possible while encouraging editors and corpus compilers to avoid or minimize the degree of editorial intervention.

2.2. Beowulf

Beowulf is an Old English poem that is revered today as one of the greatest literary monuments of Early Medieval history. However, it survives only in one known manuscript, Cotton Vitellius A. XV (henceforth, the Beowulf manuscript). The manuscript was originally part of Sir Robert Cotton's collection in Ashburnham House but was ultimately acquired by the British Library in 1753 after a fire in 1731 that resulted in the loss of innumerous manuscripts (Kiernan 1996:67-68).[2] Remarkably, Beowulf survived, but the damages of the fire are still visibly present on the manuscript, with scorched edges and many disintegrated folio margins. Before 1815, there was no edition of Beowulf that the general public could access. The only way to read the manuscript was to go to the British Library. It was not until the Danish scholar Grimm Jónsson Thorkelin transliterated the Beowulf manuscript (Thorkelin A: by a professional copyist in 1787; Thorkelin B: by Thorkelin himself ca. 1790) that an edition of Beowulf was made available (Thorkelin 1815). Since then, many edited versions and translations of Beowulf have emerged (Kemble 1833; Grundtvog 1861; Klaeber 1922; Dobbie 1953; Swanton 1978), with Klaeber's edition, now in its fourth iteration (Fulk et al. 2008), serving as one of the authoritative versions of the text. High-resolution digitized images of the Beowulf manuscript, along with the Thorkelin transcripts, are now publicly available through the Electronic Beowulf project led by Kevin Kiernan.[3] Users can find all extant folios, turn the leaves, and digitally zoom in and out to inspect the manuscript remotely, as if they were there holding the manuscript in person.

Paleographic evidence suggests that the Beowulf manuscript was copied from an Exemplar by two hands in the tenth or eleventh century (Ker 1957; Kiernan 1996), referred to as Scribe A and Scribe B (Dumville 1988). An Exemplar is a text from which another version was copied which is not necessarily the original copy, known as the archetype. The date of the archetype is debated (Chase 1981; Lapidge 2000; Neidorf 2014), with two major schools of thought: early dating (pre-750 CE) and late dating (post-750 CE). At a minimum, from the content the poem can be postdated after 531 CE based on reference to an event that is recorded in history that is mentioned in the poem (Bjork & Obermeier 1997), namely *the Frisian raid*. This, in addition to Christian symbolism, suggests the poem was composed after conversion to Christianity and after *the Frisian raid*. Given the lack of linguistic influence of Old Norse, despite being a poem set in

Scandinavia, the poem has also been argued to be pre-Viking age (Fulk 1992), that is, prior to the late eighth century. Based on differences in handwriting and script, Scribe A, who writes in Insular miniscule, appears to have copied the text up to line 1939, and Scribe B, who writes in square miniscule, appears to have copied from line 1940 onward.

2.3. Helsinki Corpus

While many historical corpora are available for studying the history of the English language (Stratton 2020a:203-205), the *Helsinki Corpus of English Texts* (Rissanen, Kytö & Palander-Collin 1993), which contains around 1.5 million words from Old English to Early Modern English, is still one of the few long-diachrony structured corpora available for the history of English. In the year 2000, Rissanen estimated that the availability of this corpus led to the publication of hundreds of diachronic analyses, a figure that must be much larger today (Rissanen 2000). The Helsinki corpus may be considered a go-to resource in diachronic analyses of English because, unlike other corpora (e.g., the Penn Corpora: Kroch & Taylor 2000; Taylor, Warner, Pintzuk & Beths 2003; Kroch, Santorini & Delfs 2004), it is currently freely available online without a license (Kroch & Taylor 2000; Kroch, Santorini & Delfs 2004). Moreover, it contains texts from various periods (Old English, Middle English, Early Modern English), and eleven subperiods of English (e.g., O1, O2, O3, O4), covering a time depth of approximately 1000 years (from ca. eighth century to the eighteenth century). The lack of complex syntactic annotation may make the corpus more user-friendly and attractive to the average corpus user, and the external metadata such as text type and genre makes the data particularly amenable to socio-historical analysis.

2.4. Analysis

2.4.1. Spacing. Beginning with the first folio (folio 129r, BL 132r), we see that some words that are written together in the edited versions of Beowulf, as well as in the Helsinki corpus (e.g., *gefrunon*, *fremedon*, *feasceaft*) are not written together in the original manuscript (e.g., *ȝe frunon*, *fre medon*, *fea sceaft*). Looking beyond the first leaf, the opposite trend can be observed too, with words written separately in the corpus (e.g., *to life*) that are written without spacing in the original (e.g., *tolife*). While the editorial decision to write words with prefixes (e.g., *ȝe*) together with the rest of the word (e.g., *frunon*), as in *gefrunon*, seems reasonable (the same being true for separating what are seemingly two distinct words, such as writing *tolife* as *to life* 'to life'), one has to wonder why the scribes explicitly decided to use spacing in this way and consider what may be lost or misrepresented in changing it. One hypothesis could be that scribes were trying to indicate morpheme boundaries (Cyrus 1971), since it was common in Early Medieval manuscripts to use spacing to indicate "morphemic hierarchical word blocks" (cf. Saenger 1997:97). However, this hypothesis would not explain why the prefixes (e.g., *ȝe*) are separated with a space, but other morphemes are not (e.g., *-on*). In work on Middle and Early Modern English manuscripts, Palmer (2009:79) observed some spatial divisions that were not indicative of morphological boundaries (e.g., *no table* for *notable*), indicating that scribes may have been using spacing to indicate syllable boundaries.

While removing spaces between words like *gefrunon* may seem minor, and editorial changes of this kind may ease modern readability of historical texts, spacing becomes somewhat more problematic with the representation of compounds. For instance, *fela* 'much, lots, very' (cognate with German *viel* 'lots') can be used as an intensifier in Old English, intensifying adjectives (e.g., *felageong* 'very young'), a quality that is true across the Early Germanic languages (e.g., Old High German *filu kuani* 'very audacious' and Old Saxon: *filo uuis* 'very wise'). However, unlike in other Early Germanic languages where the cognates of *fela* are usually considered an analytic intensifier that can stand alone, in Old English, most scholars categorize *fela* as a bound morpheme (Borst 1902:21; Ingersoll 1978:91; Lenker 2008:250; Méndez-Naya 2021). For instance, Borst (1902:21) writes that "*fela*- wird als Kompositionspartikel nur im Ae. gebraucht" '*fela*- is used as a compound particle only in Old English.' In the Helsinki corpus, as well as in edited versions of Beowulf (e.g., Fulk et al. 2008:374), *fela* is written together with the element it intensifies, mirroring other bound intensifiers which are also written together (e.g., *ærgod* 'very good,' *forlytel* 'very little,' cf. Méndez-Naya 2021). However, contrary to how it is written in historical corpora and edited editions of texts, *fela* is not written together with the element it intensifies in the original manuscripts (see Figures 1 and 2 for *fela* in Beowulf). Nevertheless, in edited versions of Beowulf, as well as in the Helsinki corpus, *fela* is written together, as in *felahror* 'very strong' and *felamodigra* 'very brave.' Spacing was likely intentional in the Beowulf manuscript, as it was in most Early Medieval manuscripts (Saenger 1997), and *fela* is consistently written with a space before the element it modifies. Parchment and vellum were hard to obtain (Clemens & Graham 2007:23), so including a space in a manuscript as opposed to using all the available space (i.e., *scriptio continua*) must be taken as intentional on the part of the scribe(s). The space between *fela* and the modified head is not a partial space unlike with *ȝe frunon*. On the contrary, the length of the space appears to be of the same length used when indicating word boundaries (the space between *sinniȝne* 'sinning' and *secȝ* 'man/warrior' in Figure 2 is not shorter than the space between *fela* 'very' and *sinnigne*). It is unclear on what grounds editors decided *fela* should be written together and whether this decision influenced the claim that *fela* is a bound morpheme in Old English. What is clear, though, is that *fela* cannot be considered a bound morpheme on the basis of an absence of spacing in the surviving manuscript.

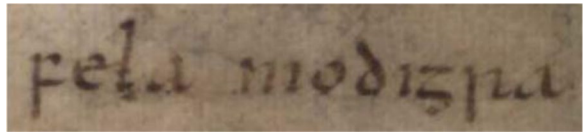


Figure 1. *Fela modiȝra* 'very brave.GEN.PL'
Source: Image of London, British Library, Cotton Vitellius A. XV. fol. 171r, reproduced from Kiernan's *Electronic Beowulf 4.0* with permission from Kevin Kiernan.

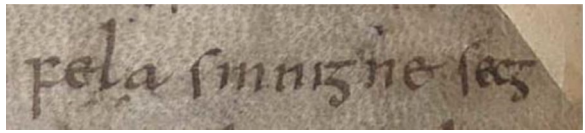


Figure 2. *Fela sinnigne secg* 'very sinful warrior'
Source: Image of London, British Library, Cotton Vitellius A. XV. fol. 160v, BL 163v, in Kiernan's *Electronic Beowulf 4.0* with permission from Kevin Kiernan.

In checking Thorkelin's transcripts, compounding cannot be attributed to Thorkelin either, since he transliterates *fela* faithfully with spaces (e.g., 176r-r). So where did the notion that the intensifier *fela* was a bound morpheme come from? It is not compounded in most other Early Germanic languages (with the exception of Old Norse edited textual editions) and there is no orthographic evidence to suggest it was compounded in Old English either.[4] Even beyond the Beowulf manuscript, the intensifier *fela* is followed by an immediate space, despite being written with its modified head as one unit in edited editions and corpora, including the Helsinki corpus (either by writing them together with no spacing or by hyphenating). For instance, although *fela* is treated as a bound morpheme in edited versions of the Anglo-Saxon rune poem, it is not written together with the element it intensifies in the original (Cotton Otho B.x, fol. 165a-165b). The same is true for the use of *fela* in Maxims I (*felameahtig* 'very mighty'), which is compounded and written without spacing in edited editions and historical corpora, but not in the original manuscript in the Exeter Book (Exeter Book, Maxims I, 90r). One possible explanation is that *fela* is treated as a compound by modern editors because of the metrical requirements of Early Germanic verse, laid out in Eduard Sievers' *Altgermanische Metrik* (1893). According to Sievers, lines in Old English verse, called *Langzeile*, consist of two short lines (*Kurzzeile*), and at least one lexical item in each short line alliterates. Perhaps writing *fela* without a space could affect Sievers' metrics, but the metrical requirements do not explain why *fela* is treated as a free-standing morpheme in other Early Germanic languages (e.g., Old High German) where the same metrical requirements are supposed to apply.[5]

The fact that the intensifier *fela* is attested modifying a wide range of heads (ca. 20-30 according to *The Dictionary of Old English*) could theoretically be taken as preliminary evidence that it was a free morpheme in Old English, since free morphemes tend to be productive (Bauer 2003; Haspelmath & Sims 2010). However, since high type frequency is not exclusive to free forms, as it may also be a characteristic of affixes and affixoids (Kim 2015), productivity alone cannot reliably indicate morphological status. In addition to functioning as an intensifier, *fela* was also used as a quantifier in partitive genitive constructions (e.g., *fela monna* 'many men'), a usage that is both more frequently attested and etymologically older. Its cognates in other Indo-European languages (e.g., Latin *plus* 'more/much'), as well as its reconstruction in Proto Germanic (*feluz), illustrate that the quantifier function predates the intensifier function. Thus, based on the semantic developmental chronology (i.e., from quantifier to intensifier), it appears *fela* was in the process of grammaticalization in Old English.

Kim (2015:429-430) outlines four types of semantic change of Old English "prefix-like morphemes." *Fela* fits into the second type—(abstracting/intensifying)—similar to *in* (*infrod* 'very old' versus *in* 'in').[6] This type represents an initial stage of grammaticalization. Kim (2015) also argues that the distinction between free and bound (specifically, free versus prefix) should be part of a continuum rather than a binary, a view shared by Booij (2005:111) who argues that "there is no sharp boundary between compounding and affixal derivation." The fact that *fela* is more frequently attested as a quantifier than an intensifier suggests the morphological status of *fela* in Old English was closer to a free morpheme on this continuum.[7] *Fela* is also attested as a modifier of adverbs (e.g., *hie fela swiðor wepan* 'they wept very intensely'), which is typically the duty of free as opposed to bound morphemes.

Returning to the issue of spacing, an interesting parallel can be found in Modern Dutch: Booij (2005:116) uses spacing as evidence that *bere* 'bear' in Dutch has been reanalyzed, by some speakers, as an intensifying adverb (e.g., *bere goed* 'very good'). This example, in line with Palmer (2009:29), suggests that linguists should pay attention to orthographic boundaries as potential evidence of speakers' perceptions of word boundaries and word status. Typographic conventions are especially relevant for historical linguists who lack the kinds of experimental or introspective data that can be collected when determining morphological status in the present day (i.e., historical linguists cannot elicit data from participants, nor can they carry out grammatical judgment tests).

While spacing may seem trivial, the editorial decision to write a word together or not can have theoretical and methodological implications. On a theoretical level, writing two elements together may lead scholars to treat the absence of a space as indicative of compounding or bound morphological status, which may have been the case for *fela*. The scribe(s) may have intentionally written words separately to indicate morphological status or to indicate syllable boundaries, an intention that is ultimately lost if changed by editors, and is only recoverable by examining the original manuscript. Spatial organization in writing also has an effect on cognitive representation (Saenger 1997), and writing an element by itself with a space could be interpreted as evidence that the scribe viewed it as an independent unit (Palmer 2009:77). Since most nouns in Beowulf that are considered compounds by modern editors are not written together in the surviving manuscript, the fact that they were not written together may suggest that scribes did not consider these to be "univerbated" (Lass 2004:35-36).[8] Moreover, given that in edited versions spaces are artificially added to verse texts (known as a caesura) to conform to Sievers' (1893) metrics, modern readers of Old English may have spatial expectations and search for visual cues to indicate whether the text is a verse text or a prose text (Carpenter 2018:18-19).[9] The Helsinki corpus, however, does not use caesuras.

The insertion of spaces can also have methodological implications for linguistic analyses. For instance, a common measurement for word frequency in corpus-based work is *frequency normalization*, which is calculated based on the number of words in a corpus (Biber, Conrad & Reppen 1998:263-264). If corpora do not accountably reflect spacing, it could impact the assumptions and conclusions that users make when quantifying word frequencies: if spaces are added or removed, frequency is affected. While it is not entirely clear whether *fela* was a free or bound morpheme in Old English, corpus users who do not check or are unable to gain access to the original

manuscripts are left to the discretion of the editors. If editors decide elements should be written together with the element it modifies, this may be taken by users as unsubstantiated evidence for a particular theory (e.g., *fela* is a bound morpheme).

2.4.2. Punctuation. Punctuation, while present in some manuscripts, was not used in the same way in edited editions and historical corpora as the original manuscripts. For instance, the first word of Beowulf, *hwæt*, which has received ample attention in the literature (e.g., Stanley 2000; Walkden 2013; Brinton 2017:45), is sometimes written with an exclamation mark, even though exclamation marks, like many other punctuation signs, did not exist at the time and thus do not appear in the manuscript. Exclamation marks do not appear until the fourteenth century (Parkes 1992). The exclamation mark after *hwæt* appears to be due to John Mitchell Kemble's 1833 edition of Beowulf, as a result of which "editors have felt free to litter the poem with exclamation points" (Weiskott 2012:27). According to Weiskott (2012:27), Grundtvog's 1861 edition of Beowulf contained eleven exclamation marks throughout the first eight hundred lines, "only to be outdone six years later by C. W. M Grein, who populated his Beowulf with a frenetic forty-five in the first five hundred lines alone." The mass insertion of anachronistic exclamation marks raises the question of whether modern-day assumptions about language are being superimposed on the historical form and whether this affects our interpretation.

Not only is punctuation added in editorial versions and corpora that was not in the original, as Mitchell (1980:25) points out, much of the punctuation in earlier editions was not based on modern English but rather on modern German, a language in which, prescriptively, punctuation (specifically commas) serves specific grammatical functions (e.g., to indicate clausal boundaries). While the Helsinki corpus does not include an exclamation mark after the opening *hwæt*, it does include a comma before *hu* 'how' that is not there in the original manuscript (see Figure 3). Inserting a comma in the corpus may superimpose the view that *hu ða æþelingas ellen fremedon* is a declarative sentence 'how the nobles achieved greatness' (i.e., we will learn how they achieved greatness). However, *hu ða æþelingas ellen fremedon* may be an exclamatory sentence (i.e., 'how they achieved such greatness!'), since PDE *how* can serve as a common exclamatory device (e.g., *how nice!*). Not putting in a comma allows the user to decide for themselves what the appropriate interpretation of the sentence is.

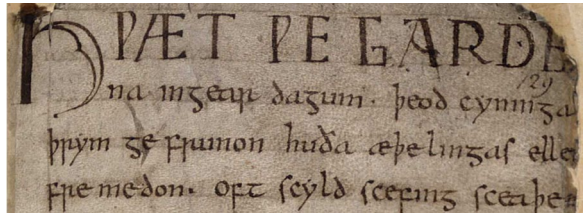


Figure 3. Opening Lines of Beowulf
Source: Image of London, British Library, Cotton Vitellius A. XV. fol. 129r, BL 132r in Kiernan's *Electronic Beowulf 4.0* with permission from Kevin Kiernan.

A close inspection of the first folio of the Beowulf manuscript reveals the presence of several dot-like markings (or Latin *punctus*), over sixteen period-like marks according to my estimates, not including the raised punctus above the letter <ẏ> which was punctuated in Anglo-Saxon miniscule.[10] While the use of the punctus in the manuscript may not map onto our contemporary uses of punctuation, puncti (plural of *punctus*) were likely still meaningful for the scribes and contemporary orators (Parkes 1978; Weiskott 2012:40). Scribes may have used such punctuation marks to encourage readers to read and interpret a text in a specific way (Parkes 1997:8). By altering punctuation to conform to modern standards, editors may be imposing a particular interpretation of the text onto readers. In edited versions of Beowulf, line 99 has a comma inserted after *eadiȝlice* which does not appear in the manuscript (folio 132r). While seemingly innocent, an anachronistic comma here imposes a particular syntactic and prosodic interpretation on the reader. While there are at least two, possibly three readings, the editorial decision to insert a comma removes ambiguity and favors only one (see Lass 2004 for discussion). The Helsinki corpus, like edited versions, also inserts an anachronistic comma after *eadiȝlice*, which could affect a syntactician's analysis when using the corpus as a source of linguistic data. As Lass (2004:25) writes, "no modern (or any) editor can be said to know the language of a scribe better than the scribe did." Since it is not possible to know what information will be useful to researchers, corpus compilers should consider including whatever devices scribes used that are visible on the manuscripts, as they could be invaluable to modern interpretations of the text, even if they are not fully understood by the editors at the time (Scragg 1971:699).

A period that appears equidistant between two words in the center, as in Figure 3 (see period between *dagum* 'day.DAT.PL' and *þeod* 'people,' is called a *punctus circumflexus*. Donoghue (2006) points out that there are approximately twenty of these periods every hundred lines in Beowulf which may facilitate scansion, that is, the measurement of long and short syllables.

> In the Beowulf manuscript, many points fall just before a recognizable metrical and syntactic pattern at the beginning of the alliterative line (Donoghue 2006:51).

Donoghue also points out that Scribe A was more deliberate with "pointing" out metrical and syntactic features than Scribe B, showing interscribal variation. While it is easy to sideline the occurrence of small diacritical elements such as a period, analyses of the punctus in medieval manuscripts have revealed important information not only about suprasegmental features such as stress, prosody, and metrics, but also the socio-historical context in which the manuscript was used. For instance, in his analysis of the punctus in two Old Saxon manuscripts of the *Heliand*, Sundquist (2019) argued that in one version of the manuscript (Monacensis BSB cgm 25, henceforth MS M), the punctus is systematic, occurring where metrical and syntactic boundaries coincide, suggesting that the manuscript was used for recitation purposes, with the periods serving as an important rhetorical aid in a type of oral performance. MS M also contains passages with musical notes and accent markings, which corroborate the idea that the manuscript was used for oral recitation. Nevertheless, early editors of the *Heliand*, the

longest attested Old Saxon record, disregarded these marks as "vollkommen willkürlich" 'completely arbitrary' (Sievers 1878). Similarly, scholars such as Lennard (1999) and Calle-Martín and Miranda-García (2005) argue that medieval punctuation can be grammatical and can perform rhetorical and elocutionary functions. For instance, in the Lambeth Psalter (MS 427), a biblical text written in Latin, Old English glosses are provided for the Latin words, but dots are used underneath the words to indicate contemporary Old English word order (Robinson 1973).[11] People using the manuscript were therefore clearly using diacritical marks to aid interpretation. As Lass (2004:35) notes, there appears to be "a certain amount of prosodic marking" in the Beowulf manuscript, which editors do not consider. In a comparative analysis of the two fragments of the *Heliand* (MS L and P), Sundquist (2024) argued that puncti can be used to examine relationships between manuscripts, with the suggestion that these two fragments likely came from the same codex.[12] While one justification for excluding diacritical marks in corpora could be that they interfere with search queries, issues with search query syntax could be resolved through layered annotation (e.g., Marttila 2014), which preserves original punctuation while allowing searches for a normalized version.

In summary, historical corpora contain punctuation that was not there in the original manuscript(s) but remove diacritical marks that were. By artificially adding punctuation, a corpus may superimpose a particular reading on the user, which in turn, may affect a linguistic analysis. Moreover, by adding punctuation to match modern conventions and expectations anachronistically assumes modern rules apply to the past. The replacement or removal of diacritical elements could also strip away the social meaning embedded in a text, while superimposing anachronistic or false interpretations.

2.4.3. Spelling. Orthographic variation is a well-known problem in historical corpus linguistics (Piotrowski 2012), with researchers having to rely either on canonical spellings when running search queries or adding potential spelling variants by utilizing, for instance, the OED, via the word list function on a concordancer (Curzan & Palmer 2006). Some tools have been designed to deal with orthographic variation by searching for potential spelling variants (Baron & Rayson 2008; Baron, Rayson & Archer 2009; Burns 2013), but it is unclear how many non-canonical spellings still go unrecognized and to what extent this affects a historical corpus-based analysis (Curzan & Palmer 2006:24). While spelling differences are still preserved in the Helsinki corpus due to the lack of lemmatization, standardizing spelling could potentially remove hidden social meaning embedded in orthographic variation. On the one hand, orthographic variation may indicate phonological differences, which in turn can point to geographical information, but it could also indicate social meaning. Just like <analyse> and <analyze> can be pronounced in the same way today but one orthographic form indexes a British identity while the other marks a North American identity, orthographic variation across scribes may also be important in reconstructing the contemporary social meaning. Nevertheless, as Sairio, Kaislaniemi, Merikallio, and Nevalainen (2018) point out, it is rare to find a text in a corpus that has not undergone some orthographic edits. In the Helsinki corpus, as well as textual editions of Beowulf,

Old English symbols that are no longer used in Modern English have been modernized (e.g., ƿ > w, ȝ > g, ꝛ > r).

When corpus compilers make the conscious decision to improve the readability of a historical text from a modern perspective by standardizing spelling, choices may not be problematic to, say, syntacticians, but they may be changing vital information that to phonologists and historical sociolinguists would be crucial (Curzan & Palmer 2006).[13] Editors should also consider the ethical implications of normalizing spelling. Does standardizing spelling send the message that the language was more homogeneous than it was? Are we removing a visible part of the identity of scribes by leveling spelling and removing variation? Scribal errors are also sometimes edited in modern editions and corpora, a practice which could be problematic for at least two reasons. First, do we truly know it was an error? Sometimes the marks on the manuscript make it clear that scribes have tried to correct errors, elements which could be important to preserve in historical corpora because they point to conscious choices the scribes made. In the *Emerging Standards Corpus*, Auer, Gordon, and Olson (2016) have made conscious choices to indicate to users when words have been added between the lines or in the margins or when scribes scribbled out a word. However, sometimes there are no indications, but editors assume there was a scribal error based on our modern understanding of the text and may alter, for instance, the grammar or spelling.[14] Second, scribal errors may be meaningful to some researchers. Some scribal errors in Beowulf, for instance, may point to its provenance. For instance, Lapidge (2000) suggests that the extant manuscript was copied from a much earlier text, prior to 750 CE in Anglo-Saxon minuscule script, with scribes confusing letters because they were copying from a slightly different script. Kiernan (1996:10) discusses the informative nature of scribal errors in Beowulf because they can be used to recreate the cause of the error. Nevertheless, "no valid assessment of the reliability of the scribes [. . .] can afford to ignore the scribes' written corrections and erasures, and yet no editor has ever taken them into account" (Kiernan 1996:10). Only by inspecting the original manuscript, or high-resolution facsimiles thereof, can we see evidence of scribes correcting their errors. For instance, on folio 136v, the scribe altered *fæft* with *fæſt*, by scraping away the cross-stroke on the *f* in the latter instance, but leaving the stroke still identifiable upon close inspection (Kiernan 1996:197). Kiernan (1996) also shows how some letters that were removed by scribes by scratching out the ink, errors that are not obvious to the naked eye, can be recovered by looking at manuscripts through ultraviolet light (e.g., the changing of *m* to *n*, which could be important for historical linguists in pinpointing phonological and morphological changes underway in Old English). While ultraviolet imaging requires access to the original manuscript, ultraviolet images can be taken and added to corpora as a separate annotation layer.

2.4.4. The Manuscript. The manuscript as an artefact can also provide socio-historical information to researchers that is not usually conveyed in a historical corpus. For instance, the poem Beowulf, despite being revered today, is written on material that suggests the text might not have been as significant in its time. While parts of the manuscript were damaged in the Ashburnham Fire and the manuscript was left uncurated for

some time in the British Museum, leading to the decay of the margins and several corners, it is evident the manuscript material was not pristine to begin with. The vellum is sturdy and durable, which helped keep the manuscript intact during the fire, but visible holes can be found on the manuscript that were clearly there at the time the scribes copied Beowulf, as scribes wrote around them (see Figures 4 and 5). Whether scribes wrote around holes or not can indicate when the holes appeared (i.e., before or after composition).

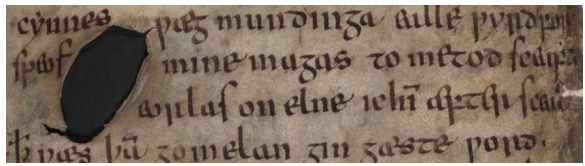


Figure 4. Example Hole in the Beowulf MS
Source: Image of London, British Library, Cotton Vitellius A. XV. fol. 191r, BL 195r in Kiernan's *Electronic Beowulf 4.0* with permission from Kevin Kiernan.

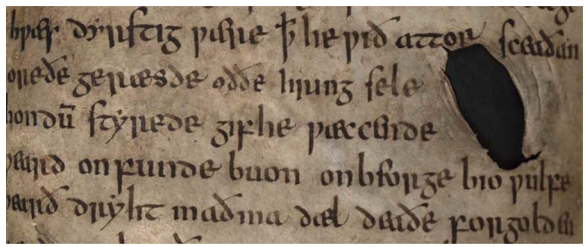


Figure 5. Another Example Hole in the Beowulf MS
Source: Image of London, British Library, Cotton Vitellius A. XV. folio 191v, BL 195v in Kiernan's *Electronic Beowulf 4.0* with permission from Kevin Kiernan.

Folio 179, upon which lines 2207-2252 are written, is also a palimpsest, that is, the ink on the parchment was scratched out to make room for something else (Kiernan 1996:11). On the one hand, this could mean the vellum was taken from some other text and was reused in second-hand style. On the other hand, Kiernan (1996:11) suggests a scribe revised the poem. Kiernan (1996:219) also suggests that the first folio of the manuscript has left traces of its former use as the outside cover of the poem, with many letters in the bottom right-hand corner being soiled and the ink diluted, caused by sweat and friction from the alleged gripping of the manuscript by the corner: "the area of the damage is restricted to the space that a thumb would occupy."

Calligraphy and handwriting, while lost and standardized in most historical corpora, can also reveal important information about the provenance of a text, the number of scribes and hands involved, dialectal variation, and the possible intention of the scribe(s).

To the latter point, Scribe A of the Beowulf manuscript writes in round miniscule whereas Scribe B writes in square miniscule. Since square miniscule is thought to be an earlier development than round miniscule, one interpretation could be that Scribe B was deliberately trying to write archaically (Kiernan 1996; see, however, Dumville 1988 and Kiernan 2021 for debate). The differences in script may also suggest that the scribes were copying from different exemplars (Voth 2017:119). This paleographical information, however, is lost in historical corpora, which usually do not indicate a change in scribal hand, such as the transition from Scribe A to B in the Beowulf manuscript. Indicating a change in scribal hand to the corpus-user could also be important for analyses of linguistic variation, since differences could be treated as intrascribal (i.e., variation within one scribe/individual) when they may in fact be attributable to interscribal variation (i.e., variation due to differences across scribes/individuals). Methodologically, not providing this information to the user could impact quantitative modeling. For instance, it is becoming standard practice in historical sociolinguistic research to include the TEXT ID as a random factor in mixed effects models (e.g., Stratton 2022, 2023) in lieu of the "speaker," which is often included in studies of modern variation. However, by including TEXT ID as a random factor, the model is assuming variation in a particular text (e.g., Beowulf) comes from one source, which is not the case in Beowulf. Interscribal information could be used as a proxy for speaker variation. If the text is annotated for scribal hand, the analyst can include scribal hand in the random factor TEXT ID as opposed to treating the text as homogenized.[15] Similarly, Scribe A is thought to have copied other prose texts in the Nowel Codex, but because this scribal information is not provided in historical corpora, including TEXT ID as a random intercept assumes there is no relationship between the contexts and that they were copied by different scribes. To paraphrase Driscoll (2010), a physical manuscript is integral to a text's meaning, and it serves as a living testimony of the hands involved in the production of the text, what has happened to it over time, and where it has been disseminated.

2.5. Recommendations for Faithfulness

Section 2 has discussed issues concerning faithfulness and how it affects our understanding of historical language. I now provide my first set of recommendations: one for the historical-corpus user and one for the corpus builder.

User:

- Language historians should check the original manuscript, or photocopies thereof, when one is available. While some manuscripts are not available online, meaning that we are left at the discretion of the corpus compilers/builders or we have to travel to archives ourselves (a fact that is currently true for those studying Gothic and Old Low Franconian), an increasing number of manuscripts are being digitized and are becoming available online, especially for Old English. When users are not able to check the originals nor do they have access to photocopies, they should think critically about whose hands were involved in the production of the corpus and whose language conventions are being accounted

for. They should inspect the historical corpus and read the corpus manual to determine what editorial changes have been made. When finding changes, they should ask themselves to what extent editorial interventions may influence the conclusions and assumptions they may make based on the text they find in the corpus. If it is the researcher's aim to analyze the language of the past, of a particular period, then it is important to ensure they are indeed analyzing the language and conventions of that period, and not a later period's reinterpretation or modernization of that language. Are they studying conventions that were found in the preserved manuscript or are they using conventions artificially transposed onto the text to fit expectations of modern readers? In other words, language historians must look under the hood of the corpus to examine whose language and conventions they are studying and be transparent about how these differences may affect their argument or conclusions.

Corpus builder/compiler:

- Check your assumptions. Who will be your audience? How might your decisions affect your users? Corpus compilers and builders may omit or change elements from the original manuscript to make texts more legible for their readership (as in printed textbook versions of Old English texts), but historical linguists may need this information. What might not be of great importance to a syntactician may be crucial for a morphologist or phonologist. Artificially inserting spaces between elements can affect a morphologist's analysis, artificially inserting commas can affect a syntactician's analysis, and normalizing spelling can affect a phonologist's and historical sociolinguist's analysis. Are you designing the corpus for a specific target audience? If so, is the target audience spelled out to the users in the corpus manual? Is any relevant textual information covered in the manual? Corpus compilers should also try to copy historical texts from the originals, when possible, as an error by a copier at one point in time can be repeated multiple times and spread. Are you editing the language in ways that resemble modern English conventions that may not reflect the actual status of the language historically? In what ways could these editorial choices affect the users' conclusions? Including different versions or layers of annotation (e.g., faithful version versus edited version) can assist in preventing and avoiding some of these issues.[16] Are users aware of your editorial changes? "Silent" editing should be avoided, and editorial changes should be made explicit to the users.

3. Does Sampling Affect Your Research Questions?

3.1. Representativeness

Most textbooks on corpus linguistics include a discussion of "representativeness," that is, the extent to which a sample in a corpus includes the full range of variability in a population (Biber 1993). Representativeness is thus an issue that corpus and historical

linguists are usually cognizant of and seek to address (Nevalainen & Raumolin-Brunberg 1989; Kytö & Pahta 2012). However, diachronic corpora cannot be sampled in the same way as corpora representing current language use (Stratton 2020a:202). Since "historical documents survive by chance and not by design" (Labov 1994:11), a historical analysis based on attested language can only be defined by the extant manuscripts. For instance, it is well known that historical English is skewed toward male upper-class literates (Hogg 2006:395; Claridge 2008:248; Auer, Gordon & Olson 2016). While historical English is better attested relative to its Germanic siblings, one cannot assume that the attested language is representative of the historical English-speaking population (Stratton 2020a:205). This section elaborates on issues concerning representativeness in historical studies based on historical corpora, discussing the ways sampling can influence conclusions about language variation and change.

3.2. The Existence Question

It is a well-known maxim in historical linguistics that the lack of attestation does not mean something did not exist (e.g., Stratton 2020a:205). Researchers interested in the existence question (i.e., did X structure exist in language Y?) must bear this fact in mind. In some instances, linguistic structures or forms may be attested but they are simply not represented in available historical corpora. The Helsinki corpus, for instance, while important for its diachronic coverage, is relatively small. There are approximately 3.5 million attested words of Old English (not unique/different words), with ca. 3000 texts (Dictionary of Old English), but the Helsinki corpus records only a fraction of that (11.8 percent). In addition to the availability issue, the language of most speakers in Early Medieval England is not attested, a reality that is true for most historical periods. Old English, for instance, is attested predominantly in the form of one regional variety, West Saxon, which makes up around 90 percent of the composite Old English corpus. The other three known remaining geographical varieties (Northumbrian, Mercian, and Kentish) are sparsely attested. In addition to the issue of regional representation, on the social side, texts from Early Medieval England are skewed toward male scribes who were likely highly educated and do not represent the speech of the everyday individual. These issues are complicated further with the fact that they are written records, which may be more formal and archaic and likely differ greatly from the spoken vernacular.

Most textbooks on Old English give the impression that a single homogenous inflectional ending for the third-person singular in the present indicative existed, namely one containing an interdental fricative (either <þ> or <ð>). However, by Middle English, there is well documented and widely studied variation between the *þ*/*-th*-ending occurring more frequently in southern English texts and *-es* occurring more frequently in northern English texts. The representation of this difference in Middle English may leave the impression that this is a Middle English phenomenon, even though this variation already existed in Old English (e.g., Blakeley 1949; Van Gelderen 2019). For instance, in the Lindisfarne Gospels (London, British Library Cotton MS Nero D.IV), which are written in Latin, some glosses are provided in Old English, indicating that speakers of Old English, likely one or several monks, were

glossing words in their vernacular to aid in the interpretation of the manuscript. A triple gloss is provided for the Latin *arbitretur* 'think'—*he lettes*, *he doemeð*, *he woenes*—indicating that the third-person inflectional ending *-es* was already present in Old English. Since the glosses were written in a Northumbrian dialect, and Old English is mostly attested in the West Saxon dialect, there is a representation problem. While the *-es* form existed, researchers cannot easily study its frequency given the lack of Northumbrian representation in the record of Old English.[17]

Representation and sampling issues are not unique to Old English. For instance, as I have shown in my own work, in modern varieties of British English, *well* can function as an intensifier of adjectives (Stratton 2018, 2020b), as in *he's well boring* and *that's well funny*. However, this use of *well* is stigmatized (Stratton 2020b), in part because it is assumed to be an innovation, particularly among adolescents (Stenström 2000). In fact, its use by adolescents may have led to its stigmatization. While *well* was used as an intensifier in Old and Middle English (Ito & Tagliamonte 2003:278; Stratton 2022), according to traditional scholarship (Mustanoja 1960:319-327), after the mid-fourteenth century *well* was thought to have disappeared as an intensifier of gradable adjectives, remaining only in certain fixed collocations (e.g., *well aware*, *well able*) that remained in the language in all varieties throughout history (see Stratton 2020b for a review). This assumption that *well* disappeared by Early Modern English (other than the limited fixed collocations) was widely accepted to be true and if researchers use historical corpora that are (1) widely used in studies of the history of English and (2) document the standard incipient varieties of English, this hypothesis is supported (Stratton 2020b). However, as I have previously illustrated in Stratton (2020b), if researchers expand their dataset to include historically marginalized and peripheral varieties of English, it becomes clear that *well* never disappeared as an intensifier of gradable adjectives, but in fact was retained in some non-prestigious varieties of English that are not well documented in historical corpora. This representation issue leads to an important question about sampling: are we falsely generalizing about language variation and change based on our sampling methods? Sampling has become a crucial component of variationist quantitative analyses, but slicing historical data into small chunks (cf. Rissanen's (1989) "mystery of vanishing reliability") can exacerbate the "bad data" problem further. Consequently, Lauersdorf (2018a:112, 2018b:211-213, 2024:338-340) calls for the use of "all the data" in a historical analysis, a principle that, if not followed, can affect interpretations and conclusions (Stratton 2020b; Fuchs, 2025).

3.3. Supplementing Corpora

Although logistically it is not always possible to include all language data in a historical corpus, researchers should consider how the sample may affect the research question they are investigating. On the one hand, this question could relate to the size of a corpus. The Helsinki corpus contains around 400,000 words, but there are over 3.5 million attested words for Old English. When should researchers use all the data? Lexical items that are less frequent require much larger corpora (Nevalainen & Raumolin-Brunberg 1989:67) but smaller corpora may be more appropriate for more frequent phenomena. For example, small corpora may be appropriate for investigations of

inflectional morphology, but larger corpora may be necessary for the investigation of less-frequent derivational affixes (Curzan & Palmer 2006). Work within the variationist paradigm that requires careful qualitative circumscription of the variable context may find over 3 million words unworkable and instead turn to smaller corpora like the Helsinki Corpus (Stratton 2022, 2023). In previous work (Stratton 2020b), to investigate whether *well* was retained in varieties of English beyond the mainstream standard incipient varieties recorded in traditional corpora, I turned to dialectal records, poems, and audio recordings (including but not limited to songs, television shows, and interviews). Combining qualitative and quantitative methods was crucial given that mainstream machine-readable corpora were not available for the varieties in which *well* as an intensifier of most gradable adjectives was ultimately retained.

Beyond size, whether to supplement a corpus or not can depend on the socio-historical nature of the research question being investigated. Is it important for the research question to include speakers from different regions and social backgrounds (assuming the data are available)? For instance, Auer, Gordon, and Olson (2016) have been challenging the traditional view of the standardization of the English language by studying communities and varieties that are not well documented in historical corpora. To what extent could a researcher's question be affected by the lack of social and demographic representation in the corpus? To what extent may register, text type, and genre affect a researcher's question and how are these represented in the corpus?

Another problem that can arise in using historical corpora for linguistic analyses is that they usually use a single version of a text in situations where there are multiple.[18] When there are several variants of the same text written by different scribes (e.g., Cædmon's Hymn), the version that gets selected could influence and shape a researcher's conclusions.[19] Different versions of manuscripts can show differing levels of variation (Lass 2004; Curzan & Palmer 2006; Grund 2006) and in selecting one version over the other, we may be favoring one variety over another and also skewing the picture of variation. For instance, Curzan and Palmer (2006) highlight how the use of different versions of the same historical text can result in very different syntactic, semantic, and phonological analyses. If manuscripts show variation in orthography, choosing a manuscript version that favors one over the other might seem inconsequential but "small orthographic differences could have major significance for a historical phonologist looking at diphthongs in such texts" (Curzan & Palmer 2006:27). Similarly, Lass (2004) discusses how the different versions of Cædmon's Hymn show grammatical differences with respect to pro-drop, with earlier versions not including the subject pronoun *we*, but later versions including it. However, as Grund (2006) points out, including multiple versions in a corpus could affect frequency counts if the same part is repeated multiple times across the manuscripts, leaving the impression that a variant occurred multiple times when it may have only occurred once. One solution could be to structure corpora in a way that allows the user to calculate the number of words in their "virtual" corpus by selecting the version of the text of their choosing, and the counts are automatically updated based on selection. While including only one version of a text in a historical corpus can have theoretical and methodological implications, edited versions of historical texts can be much more problematic with respect

to variation, as some editions combine features of different versions of manuscripts (Lass 2004; Grund 2006) like a fictional Frankenstein monster that never existed, creating a false depiction of historical reality. Conflated Frankenstein texts are a reminder that edited versions of texts can impact historical linguistic analyses (e.g., Grund, Kytö & Rissanen 2004; Lass 2004), and as useful as corpora are, it is crucial to check that any assumptions or conclusions we make about historical language have not been shaped by editorial changes.

3.4. Recommendations for Sampling

Section 3 has discussed how sampling can affect a research question, discussing issues of sampling and representativeness, when it is appropriate to supplement, and which factors might influence the decision to use one historical corpus over another. I now arrive at my second set of recommendations for corpus users and corpus compilers.

User:

- Expand your dataset, combining qualitative and quantitative approaches. In questions of existence, use all the data at your disposal, supplementing historical corpora as needed. The choice of corpus size, however, is dependent on the research question. How might the sample in the corpus affect the research question? Do you need socio-historical metadata to adequately address the research question? If not, to what extent may conclusions be affected or weakened by not including the fullest picture of variation? If the data are skewed toward a particular population or group of speakers, acknowledge this issue as a methodological shortcoming when drawing conclusions, reasserting that the historical corpus data are just a sample. Hopefully it is "representative" (e.g., stratified sampling methods), but this can never be fully the case for historical data.

Corpus builder/compiler:

- Aim to create the broadest picture of variation and avoid a Frankenstein monster. When multiple versions of the same manuscript exist, consider including all versions. The different versions can appear parallel to one another or could be accessible by links and tabs, with the option to flick from version to version. Perhaps it is possible to structure the corpus in a way that allows the user to select the texts they want to use and provide them with automated word counts required for frequency normalization. Never make a Frankenstein text! Doing so is a disservice to language historians and gaslights the people whose language we are studying.

4. Conclusion

The advent of machine-readable corpora has transformed the way historical analyses can be conducted, providing a wealth of material and remote access to users with the click of a button. However, analyses based on historical corpora must be interpreted

within the context of their constraints and shortcomings. With the increasing emphasis on "big data" and statistical analyses in studies of language history, philological knowledge is lost, which in turn affects the interpretation of historical corpus-based studies. This paper discussed two salient issues that researchers and corpus compilers must address when using and building historical corpora: faithfulness and sampling.

Given that minor editorial interventions can shape and influence views about historical language, this paper calls for researchers to return to the original manuscripts when digital editions are available and advocates for minimal editing or layered annotation on the part of the editor. Faithfulness is particularly problematic for the earlier periods of the English language because researchers may not have access to the original manuscripts, nor may they have the training to check whether the information provided in the historical corpus is faithful. Non-specialists are therefore relying on the editors to provide as authentic a picture of historical language as possible. Many historical corpora use secondary or edited editions as their source texts, which many linguists and language historians use to make claims about historical language. Depending on the faithfulness of the corpus and the analyst's research questions, claims may be affected. The conclusion of Lass (2004:46) sums up this call well:

> The ideal model for a corpus or any presentation of a historical text is an archaeological site or a crime-scene: no contamination, explicit stratigraphy, and an immaculately preserved chain of custody.

Since studies of linguistic variation and change can be influenced by editorial intervention, it is of moral and ethical responsibility to ensure that information is provided to users as faithfully as possible. Presenting data in a way that has been edited paints an ahistorical version that can influence perceptions and conclusions. A seemingly innocent comma inserted for our modern eyes could affect a syntactic analysis, while including linguistic elements could affect a morphological analysis. Normalizing spelling variation may improve modern readability, which can facilitate linguistic annotation and corpus queries, but could affect phonological and historical sociolinguistic analyses. Furthermore, not including information about the scribal hand or scribal hands involved in a manuscript's production could affect analyses of variation. Editorial changes can affect all branches of linguistic analysis. If editors and corpus builders must make modifications (be it through removal of information, addition of new information, or altering existing information), editing should not be done silently. Editorial changes should be conveyed explicitly to the user.

While it has become standard practice in sociolinguistics to sample data, historical data cannot be sampled in the same way as modern language data. Language historians, especially those working on early historical periods, are left with the data that remain, which are never representative of the broader population of speakers. While no historical corpus is expected to contain all historical language data, and new manuscripts will continue to be discovered over time, a corpus compiler's selection of language material can inadvertently marginalize specific speech communities and dialects, which in turn can influence research conclusions and paint an unbalanced picture of linguistic

variation and change. Moreover, in cases where multiple manuscripts are available, the favoring of one manuscript over another, whether intentional or not, can shape the picture of variation in historical corpus research. In addition to calling for researchers to read the original manuscripts, when possible, this paper calls for researchers to expand their dataset beyond traditional corpora, including non-traditional genres and texts, written by underrepresented individuals, combining qualitative and quantitative research methods. Editors should aim to provide the broadest picture of variation possible.

Edit less and include more!

Acknowledgments

The author would like to thank the two anonymous reviewers and the editors of this special issue. The author is particularly grateful to Chris Palmer for generously devoting his time to help refine the ideas presented in this paper. Gratitude is also extended to Laurel Brinton for her comments on an earlier draft.

Declaration of Conflicting Interests

The author declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author received no financial support for the research, authorship, and/or publication of this article.

ORCID iD

James M. Stratton https://orcid.org/0000-0002-7255-5545

Notes

1. *Edited edition* is an umbrella term for a text that reproduces the contents of a manuscript with some modification (e.g., spelling, grammar, punctuation, layout).
2. Cotton Vitellius A. XV consists of multiple texts, the story of Beowulf being only one. Robert Cotton artificially combined two unrelated codices, a twelfth-century collection known as the Southwick Codex and the Nowel Codex, which contains Beowulf (Kiernan 1996:7).
3. https://ebeowulf.uky.edu/
4. In Old Saxon edited editions, *filu* is written together with *wis* (*filuwîs* 'very wise,' cf. Cathey 2000:56, 2002:44), but in the original manuscripts, it is not written together. See MS M (see Monacensis BSB cgm 25) https://www.wdl.org/en/item/4107/view/1/18/ (image 9).
5. According to Sievers, there are five metrical types for short lines (A, B, C, D, E). The first short line on line 27 in edited editions of Beowulf (*felahror feran*) appears to be type A of Sievers' metrics.
6. Note, however, that *infrod* is actually *in frod* in the Beowulf manuscript. The *in-* in *infrodum* 'very wise.DAT.SINGULAR/PLURAL' actually appears on a different leaf.

7. In his *Old English Grammar*, Wright (1908:282) also lists intensifying *fela* as an adverb, suggesting he viewed it as free, not bound.
8. Univerbation refers to the process by which single words merge into one.
9. Caesuras are artificial in the sense that the spaces do not actually exist in the manuscripts, but it is assumed by most literary scholars that the caesuras/pauses did exist in oral performance.
10. Note that, unlike in Modern English, the <i> in Old English is not dotted, but the <y> is.
11. For more discussion and for example images, see: https://thijsporck.com/2017/10/30/reading-between-the-lines/
12. Fragment L is stored at the University of Leipzig library and fragment P is stored at the University of Berlin library (although it was originally stored at the University of Prague library, hence P).
13. Because of ubiquitous orthographic editing, orthographic variation has been under-investigated in historical corpus research. However, Sairio, Kaislaniemi, Merikallio, and Nevalainen (2018) built a tool that allows analysts to check the orthographic reliability of edition-based corpora.
14. Fleischman (2000:38) discusses how an editor changed grammatical case endings in an edited version of an Old French manuscript to conform to how the editor thought cases should have been used in the eleventh century.
15. When a text was composed by multiple hands, including TEXT ID as a random factor is still important: the variation may not come from a single source but it is still constrained than if two completely different texts by different scribes were subsumed under a single ID.
16. For example, the Electronic Text Edition of Depositions provides access to both original material and a searchable interface (cf. Kytö, Grund & Walker 2011).
17. Related to section 2 on the importance of returning to the manuscript, the glosses in the Lindisfarne Gospels are traditionally thought to have been written by a single hand, Aldred, based on a colophon at the end (folio 259r). However, paleographical and philological evidence suggests it was not Aldred, who simply tries to take the credit (Cole 2016). Knowledge of this kind is not only useful for authorship but also for understanding how widely adopted the *-es* form was in Old English. If one adopts the view that Aldred wrote the glosses, the occurrence of *-es* could be attributable to an idiosyncrasy of one individual. Yet, if the Lindisfarne glosses were not written by one scribe but instead were a collective effort from Northumbrian scribes, then they would suggest the *-es* variant was more widespread across Northumbrian speech communities. A reviewer asked about the type of evidence provided that there was more than one scribe. Cole (2016) discusses various lines of evidence, such as the interpretation of the meaning of the colophon, the fact that colophons were unusual for glosses, use of different ink color for different glosses, orthographic and paleographical peculiarities, and linguistic inconsistencies across glosses.
18. Note, however, that the *Toronto Dictionary of Old English Corpus* includes multiple manuscripts of the same texts.
19. There are twenty-one copies of Cædmon's Hymn known to date, but two have been destroyed and three badly damaged (cf. O'Donnell 2018). Cotton Otho B. XI was destroyed in the same fire that damaged the Beowulf manuscript (i.e., Cotton Vitellius A. XV). The Helsinki corpus contains only one version, marked as "Northumbrian version," but there are five Northumbrian versions known to date: MS Kk.v.16, MS Lat. Q.v.I.18, MS 575, MS Hatton 43, Brussels MS 8245-57. It is a diplomatic version of MS Kk. v.16 that appears in the Helsinki corpus. A digitized version of this manuscript can be found through the University of Cambridge Digital Library (http://cudl.lib.cam.ac.uk/view/MS-KK-00005-00016/264), leaf 128v.

References

Alexander, Michael. 2005. *Beowulf: A glossed text*. London: Penguin Books.

Auer, Anita, Moragh Gordon & Mike Olson. 2016. English urban vernaculars, 1400–1700: Digitizing text from manuscript. In María José López-Couso, Belén Méndez-Naya, Paloma Núñez-Pertejo & Ignacio M. Palacios-Martínez (eds.), *Corpus linguistics on the move*, 19-40. Leiden/Boston: Brill.

Baron, Alistair & Paul Rayson. 2008. VARD 2: A tool for dealing with spelling variation in historical corpora. In *Proceedings of the postgraduate conference in corpus linguistics*. Birmingham: Aston University. https://ucrel.lancs.ac.uk/people/paul/publications/BaronRaysonAston2008.pdf

Baron, Alistair, Paul Rayson & Dawn Archer. 2009. Word frequency and key word statistics in corpus linguistics. *Anglistik* 20(1). 41-67.

Bauer, Laurie. 2003. *Introducing linguistic morphology*. 2nd edn. Edinburgh: Edinburgh University Press.

Biber, Douglas. 1993. Representativeness in corpus design. *Literary and Linguistic Computing* 8(4). 243-257.

Biber, Douglas, Susan Conrad & Randi Reppen. 1998. *Corpus linguistics: Investigating language structure and use*. Cambridge: Cambridge University Press.

Bjork, Robert E., & Anita Obermeier. 1997. Date, provenance, author, audiences. In Robert E. Bjork & John D. Niles (eds.), *A Beowulf handbook*, 13-34. Lincoln, NE: University of Nebraska Press.

Blakeley, Lesley 1949. The Lindisfarne s/ð problem. *Studia Neophilologica* 22. 15-47.

Booij, Geert. 2005. Compounding and derivation: Evidence for construction morphology. In Wolfgang U. Dressler, Dieter Kastovsky, Oskar Pfeiffer & Franz Rainer (eds.), *Morphology and its demarcations: Selected papers from the 11th morphology meeting, Vienna, February* 2004, 109-132. Amsterdam/Philadelphia: John Benjamins.

Borst, Eugen. 1902. *Die Gradadverbien im Englischen* (Anglistische Forschung 10). Heidelberg: Carl Winter.

Brinton, Laurel J. 2017. *The evolution of pragmatic markers*. Cambridge: Cambridge University Press.

Burns, Philip R. 2013. *MorphAdorner v2: A Java library for the morphological adornment of English language texts*. Evanston, IL: Northwestern University. https://morphadorner.northwestern.edu/documentation/citation/ (2023).

Calle-Martín, Javier & Antonio Miranda-García. 2005. Aspects of punctuation in the Old English Apollonius of Tyre. *Folia Linguistica Historica* 39. 95-113.

Carpenter, Leslie. 2018. *Pointing rhythm and rhyme: The role of manuscript punctuation in English literary form, ca. 1000-1300*. New York: Fordham University Doctoral dissertation.

Cathey, James. 2000. *Old Saxon*. Munich: Lincom Europa.

Cathey, James. 2002. *Hêliand: Text and commentary*. Morgantown, MV: West Virginia University Press.

Chase, Colin (ed.). 1981. *The dating of Beowulf*. Toronto, ON: University of Toronto Press.

Claridge, Claudia. 2008. Historical corpora. In Anke Lüdeling & Merja Kytö (eds.), *Corpus linguistics: An international handbook*, vol. 1, 242-258. Berlin: Walter de Gruyter.

Clemens, Raymond & Timothy Graham. 2007. *Introduction to manuscript studies*. Ithaca: Cornell University Press.

Cole, Marcelle. 2016. Identifying the author (s) of the Lindisfarne Gloss: Linguistic variation as a diagnostic for determining authorship. In Julia Fernández Cuesta & Sara M. Pons-Sanz (eds.), *The Old English glosses to the Lindisfarne Gospels: Language, author and context*, 169-188. Berlin/Boston: De Gruyter.

Curzan, Anne & Chris C. Palmer. 2006. The importance of historical corpora, reliability, and reading. In Roberta Facchinetti & Matti Rissanen (eds.), *Corpus-based studies of diachronic English*, 17-34. Lausanne: Peter Lang.

Cyrus, Virginia J. 1971. Linguistic features of scribal spacing. *Visible Language* 5(2). 101-110.

Dobbie, E. van Kirk (ed.). 1953. *Beowulf and Judith, The Anglo-Saxon poetic records: A collective edition*. New York, NY: Columbia University Press.

DOE = Dictionary of Old English: A to Le online. 2024. Angus Cameron, Ashley Crandell Amos, Antonette diPaolo Healey Roy Liuzza, Haruko Momma, Robert Getz & Stephen Pelle. (eds.). Toronto, ON: Dictionary of Old English Project.

Donoghue, Daniel. 2006. A point well taken: Manuscript punctuation and Old English poems. In John Walmsley (ed.), *Inside Old English: Essays in honour of Bruce Mitchell*, 38-58. Malden, MA: Blackwell.

Driscoll, Matthew. 2010. The words on the page: Thoughts on philology, old and new. In Judy Quinn & Emily Lethbridge (eds.), *Creating the Medieval Saga: Versions, variability, and editorial interpretations of Old Norse Saga literature*, 85-102. Odense: University Press of Southern Denmark.

Dumville, David N. 1988. Beowulf come lately: Some notes on the paleography of the Nowell-Codex. *Archiv fuer das Studium der neueren Sprachen und Literaturen* 225(1). 49-63.

Fleischman, Suzanne. 2000. Methodologies and ideologies in historical linguistics: On working with older languages. In Susan Herring, Pieter van Reenen & Lene Schløsler (eds.), *Textual parameters in older languages*, 33-58. Amsterdam/Philadelphia: John Benjamins.

Fuchs, Katrin. 2025. Socio-historical data and the need for representative historical corpora. In James M. Stratton & Karen V. Beaman (eds.), *Expanding variationist sociolinguistic research in varieties of German*, 229-248. Routledge.

Fulk, Robert D. 1992. *A history of Old English meter*. Philadelphia, PA: University of Pennsylvania Press.

Fulk, Robert D., Robert E. Bjork & John D. Niles (eds.). 2008. *Klæber's Beowulf* (4th edition). Toronto; Buffalo; London: University of Toronto Press.

Grund, Peter. 2006. Manuscripts as sources for linguistic research: A methodological case study based on the Mirror of Lights. *Journal of English Linguistics* 34(2). 105-125.

Grund, Peter, Merja Kytö & Matti Rissanen. 2004. Editing the Salem witchcraft records: An exploration of a linguistic treasury. *American Speech* 79(2). 146-167.

Grundtvog, Nicolai, F. S. 1861. *Beowulfes beorh: eller, Bjovulfs-drapen, det old-angelske heltedigt*. Copenhagen: Karl Schönbergs Forlag.

Haspelmath, Martin & Andrea D. Sims. 2010. *Understanding morphology*. 2nd edn. London: Routledge.

Hogg, Richard. 2006. Old English dialectology. In Ans van Kemenade & Bettelou Los (eds.), *Handbook of the history of English*, 395-416. Oxford: Blackwell.

Horobin, Simon. 2012. Editing Early English texts. In Terttu Nevalainen & Elizabeth Traugott (eds.), *The Oxford handbook of the history of English*, 53-62. Oxford: Oxford University Press.

Ingersoll, Sheila. M. 1978. *Intensive and restrictive modification in Old English*. California: Carl Winter.

Ito, Rika & Sali A. Tagliamonte. 2003. *Well* weird, *right* dodgy, *very* strange, *really* cool: Layering and recycling in English intensifiers. *Language in Society* 32(2). 257-279.

Kemble, John M. 1833. *The Anglo-Saxon poems of Beowulf, The travellers song and the battle of Finnesburh*. London: William Pickering.

Ker, Neil R. 1957. *Catalogue of manuscripts containing Anglo-Saxon*. Oxford: Clarendon Press.

Kiernan, Kevin S. 1996. *Beowulf and the Beowulf manuscript*. Ann Arbor, MI: University of Michigan Press.

Kiernan, Kevin S. 2021. Square miniscule in the age of Cnut the Great. *Manuscript Studies: A Journal of the Schoenberg Institute for Manuscript Studies* 6(1). 33-73.

Kim, Yookang. 2015. Demarcation of compounding and prefixation in Old English. *Linguistic Research* 32(2). 419-450.

Klaeber, Friedrich. 1922. *Beowulf and the fight at Finnsburg*. Boston/New York: D. C. Heath & Co.

Kroch, Anthony, Beatrice Santorini & Lauren Delfs. 2004. *The Penn-Helsinki Parsed Corpus of Early Modern English* (PPCEME). CD-ROM, 1st edn, release 3. Department of Linguistics, University of Pennsylvania.

Kroch, Anthony & Ann Taylor. 2000. *The Penn-Helsinki Parsed Corpus of Middle English* (PPCME2). CD-ROM, 2nd edn, release 4. Department of Linguistics, University of Pennsylvania.

Kytö, Merja. 1991. *Manual to the diachronic part of the Helsinki Corpus of English Texts: Coding conventions and lists of source texts*. Helsinki: Department of English, University of Helsinki.

Kytö, Merja, Peter Grund & Terry Walker. 2011. *Testifying to language and life in Early Modern England*. Amsterdam/Philadelphia: John Benjamins.

Kytö, Merj & Päivi Pahta. 2012. Evidence from historical corpora up to the twentieth century. In Terttu Nevalainen & Elizabeth Traugott (eds.), *The Oxford handbook of the history of English*, 123-133. Oxford: Oxford University Press.

Labov, William. 1994. *Principles of linguistic change: Internal factors*, vol. 1, *Internal factors*. Oxford: Blackwell.

Lapidge, Michael. 2000. The archetype of Beowulf. *Anglo-Saxon England* 29. 5-41.

Lass, Roger. 2004. Ut custodiant litteras: Editions, corpora and witnesshood. *Linguistic Insights - Studies in Language and Communication* 16. 21-48.

Lauersdorf, Mark R. 2018a. Linguistic visualizations as objets d'art? In Noah Bubenhofer & Marc Kupietz (eds.), *Visualisierung sprachlicher Daten* [Visualization of Linguistic Data], 91-122. Heidelberg: Heidelberg University Publishing.

Lauersdorf, Mark R. 2018b. Historical (standard) language development and the writing of historical identities: A plaidoyer for data-driven approach to the investigation of the socio-linguistic history of (not only) Slovak. In Stephen M. Dickey & Mark R. Lauersdorf (eds.), *V zeleni drželi zeleni breg* [In the Green Country, a Green Hillside]: *Studies in Honor of Marc L. Greenberg*, 199-218. Bloomington, IN: Slavica Publishers.

Lauersdorf, Mark R. 2024. In search of patterns of historical language variation and user interaction (or: Who used what linguistic features with whom, when, where, why, and how?). *JAZYKOVEDNÝ ČASOPIS* 75(3). 330-347.

Lenker, Ursula. 2008. Booster prefixes in Old English – An alternative view of the roots of ME forsooth. *English Language & Linguistics* 12(2). 245-265.

Lennard, John. 1999. Punctuation and pragmatics. In Andreas H. Jucker (ed.), *Historical pragmatics. Pragmatic developments in the history of English*, 65-98. Amsterdam/Philadelphia: John Benjamins.

Marttila, Ville. 2014. Creating digital editions for corpus linguistics: The case of Potage Dyvers, a family of six Middle English recipe collections. Helsinki: University of Helsinki Doctoral dissertation.

Méndez-Naya, Bélen. 2021. Synthetic intensification devices in Old English. *Journal of English Linguistics* 49(2). 208-227.

Mitchell, Bruce. 1980. The dangers of disguise: Old English texts in modern punctuation. *The Review of English Studies* 31(124). 385-413.

Mustanoja, Tauno. 1960. *A Middle English syntax*. Helsinki: Société Néophilologique.

Neidorf, Leonard (ed.). 2014. *The dating of Beowulf: A reassessment*. Cambridge: Boydell & Brewer Ltd.

Nevalainen, Terttu & Helena Raumolin-Brunberg. 1989. A corpus of Early Modern Standard English in a socio-historical perspective. *Neuphilologische Mitteilungen* 90(1). 67-110.

O'Donnell, Paul. 2018. *Cædmon's hymn: A multimedia study, edition, and archive internet edition*. Version 1.1 Beta 5 (SEENET Pre-Release). https://caedmon.seenet.org/index.html (2023).

Palmer, Chris C. 2009. *Borrowings, derivational morphology, and perceived productivity in English*, *1300-1600*. Ann Arbor, MI: University of Michigan Doctoral dissertation.

Parkes, Malcom B. 1978. Punctuation, or pause and effect. In James J. Murphy (ed.). *Medieval eloquence: Studies in the theory and practice of medieval rhetoric*, 138-139. Berkeley, CA: University of California Press.

Parkes, Malcolm B. 1992. *Pause and effect. An introduction to the history of punctuation in the West*. Berkley, CA: University of California Press.

Parkes, Malcom B. 1997. Punctuation in copies of Nicholas Love's Mirror of the blessed life of Jesus Christ. In P. R. Robinson & Rivkah Zim (eds.), *Pages from the past: Medieval writing skills and manuscript books*, 47-59. Routledge.

Piotrowski, Michael. 2012. *Natural language processing for historical texts. Synthesis lectures on human language technologies*. Toronto, ON: Morgan & Claypool.

Rissanen, Matti. 1989. Three problems connected with the use of diachronic corpora. *ICAME Journal* 42(1). 9-12.

Rissanen, Matti. 2000. The world of English historical corpora: From Cædmon to the computer age. *Journal of English Linguistics* 28(1). 7-20.

Rissanen, Matti, Merja Kytö & Minna Palander-Collin. 1993. *Early English in the computer age: Explorations through the Helsinki Corpus*. Berlin/New York: Mouton de Gruyter.

Robinson, Fred C. 1973. Syntactical glosses in Latin manuscripts of Anglo-Saxon provenance. *Speculum* 48(3). 443-475.

Saenger, Paul. 1997. *Space between words: The origins of silent reading*. Stanford, CA: Stanford University Press.

Sairio, Anni, Samuli Kaislaniemi, Anna Maria Merikallio & Terttu Nevalainen. 2018. Charting orthographical reliability in a corpus of English historical letters. *ICAME Journal* 42. 79-96.

Scragg, Donald G. 1971. Accent marks in the Old English Vercelli Book. *Neuphilologische Mitteilungen* 72(4). 699-710.

Sievers, Eduard. (ed.). 1878. *Heliand herausgegeben von Eduard Sievers*. Halle: Verlag der Buchhandlung des Waisenhauses.

Sievers, Eduard. 1893. *Altgermanische Metrik*. Halle: Niemeyer.

Stanley, Eric. 2000. HWAET. In Jane Roberts & Lynne Grundy (eds.), *Essays on Anglo Saxon and related themes*, 525-526. London: Kings College London.

Stenström, Anna-Brita. 2000. It's funny enough, man: Intensifiers in teenager talk. In John M. Kirk (ed.), *Corpora galore: Analyses and techniques in describing English*, 177-190. Amsterdam: Rodopi.

Stratton, James M. 2018. The use of the adjective intensifier *well* in British English: A case study of The Inbetweeners. *English Studies* 99(8). 793-816.
Stratton, James M. 2020a. Corpora and diachronic analysis of English. In Eric Friginal & Jack A. Hardy (eds.), *The Routledge handbook of corpus approaches to discourse analysis*, 202-218. London/New York: Routledge.
Stratton, James M. 2020b. A diachronic analysis of the adjective intensifier *well* from Early Modern English to Present Day English. *Canadian Journal of Linguistics* 65(2). 216-245.
Stratton, James M. 2022. Old English intensifiers. The beginnings of the English intensifier system. *Journal of Historical Linguistics* 12(1). 31-69.
Stratton, James M. 2023. Where did *wer* go? Lexical variation and change in third-person male adult noun referents in Old and Middle English. *Language Variation and Change* 35(2). 199-221.
Sundquist, John D. 2019. What's the point? Syntax, meter, and punctuation in the Old Saxon *Hêliand. Beiträge zur Geschichte der deutschen Sprache und Literatur* 41(4). 449-476.
Sundquist, John D. 2024. Meter, syntax, and the use of punctuation in the Leipzig fragment of the *Hêliand*. In Jennifer Hendriks & Richard Page (eds.), *Investigating West Germanic languages: Studies in honor of Robert B. Howell*, 32-51. Amsterdam; Philadelphia: John Benjamins.
Swanton, Michael. 1978. *Beowulf*. New York, NY: Barnes & Noble Books.
Taylor, Ann, Anthony Warner, Susan Pintzuk & Frank Beths. 2003. The York-Toronto-Helsinki Parsed Corpus of Old English Prose. Manual. http://www-users.york.ac.uk/~lang22/YCOE/YcoeHome.htm (2023).
Thorkelin, Grímur Jónsson. 1815. *De Danorum rebus gestis seculi III & IV. Poëma Danicum dialecto Anglo-Saxonica*. Havniae: Typis T.E. Rangel.
Van Gelderen, Elly. 2019. The Northumbrian Old English glosses. *NOWELE* 72(2). 119-133.
Voth, Christine. 2017. Irish pilgrims, Welsh manuscripts, and Anglo-Saxon monasteries: Was script change in tenth-century England a legacy of the Celtic world? In Mary Clayton, Alice Jorgensen & Juliet Mullins (eds.), *England, Ireland, and the insular world: Textual and material connections in the early middle ages*, 115-136. Tempe, AZ: ACMRS.
Walkden, George. 2013. The status of *hwæt* in Old English. *English Language & Linguistics* 17(3). 465-488.
Weiskott, Eric. 2012. Making Beowulf scream: Exclamation and the punctuation of Old English poetry. *The Journal of English and Germanic Philology* 111(1). 25-41.
Wright, Joseph. 1908. *Old English grammar*. London/New York and Toronto: Oxford University Press.

Author Biography

James M. Stratton is a specialist in Germanic linguistics at Pennsylvania State University. He works on language variation and change in Germanic varieties, both past and present, with a particular emphasis on lexis and discourse.