contain links to the scripts used, but to ensure reproducibility and replicability it might be useful to develop shared practices in the field. English corpus linguistics, with its long tradition of making data available, might be at the forefront of introducing such practices.

When moving beyond the contents of this individual volume to the broader theme of using social media data in English linguistics, it is clear that social media data have great potential in the evidence-based study of English (including corpus linguistics, sociolinguistics, and language variation and change). I would suggest, however, that social media data have characteristics that call for completely novel approaches, and relying on a single set of, albeit solid, methods provides insufficient results and an incomplete picture of the phenomena under study. The editors of this volume also point out, somewhat modestly, that there is an increasing need to engage in more interdisciplinary research in the study of social media. A critical reader might argue that, to understand and fully benefit from very large and often rich social media in corpus linguistics in the future, a mere integration of quantitative and qualitative alone is insufficient. The sheer size and complexity of data (both user-generated textual data and metadata) present us with a research setting that calls for transdisciplinary approaches, and also highlights the need to broaden the expertise in computational and algorithmic directions. Corpus linguists, with long traditions in combining methods, are ideally positioned to engage in fuller collaboration with, for instance, researchers in AI, computational linguists, visualization experts and data mining specialists. The present volume is a good start in that direction, but we still need a fuller integration of methods and competencies in the future.

Reviewer's address:
School of Humanities, Foreign Languages and Translation Studies
University of Eastern Finland
PO Box 111
FI-80101 Joensuu
Finland
mikko.laitinen@uef.fi

(Received 7 April 2022)

doi:10.1017/S1360674322000260

Merja Stenroos and **Kjetil V. Thengs** (eds.), *Records of real people: Linguistic variation in Middle English local documents* (Advances in Historical Sociolinguistics 11). Amsterdam: John Benjamins, 2020. Pp. viii + 310. ISBN 9789027207951.

Reviewed by Jacob Thaisen, University of Oslo

Part I of *Records of real people: Linguistic variation in Middle English local documents* lays out the Middle English Scribal Text (MEST) programme's theoretical stance. It occupies some 100 pages comprising four chapters variously authored by Merja Stenroos (chapters 1, 3, 4), Kjetil Thengs (1, 3, 4), Martti Mäkinen (2) and Geir Bergstrøm (3). The programme's stance aligns with the sociolinguistic and pragmatic turn in linguistic studies of historical English and may be labelled 'sociopragmaphilological' (p. 6), although the volume surprisingly uses this term nowhere else. The two principal tenets are that '(a) the study of early historical linguistic variation should, as far as possible, take into account both the individual text, with its textual and historical context, and the entire corpus available' (p. 6), and '(b) the material should be studied on its own terms: research questions and categorisation should reflect the characteristics of the material' (p. 6). The second tenet responds directly to William Labov's 'bad data' problem: historical materials provide 'good data' provided researchers seek to correlate attested linguistic variation with known factors or, more precisely, factors whose values are known for those materials (cf. p. 34). The historical materials to which the volume applies these principles are the Middle English Local Documents (MELD) corpus comprised exclusively of documents dating from the period 1399–1525 and specialised materials collected by the MEST programme's doctoral candidates for their theses.

The MELD corpus, downloadable from the programme's University of Stavanger web pages, comes in four flavours ranging from 'base' and 'diplomatic', which both are graphemic transcriptions with graphetic elements incorporated, to 'readable' and 'concordancer-friendly' with most abbreviations expanded and notes removed. The former two flavours retain punctuation marks, insofar as the set of marks to distinguish has been firmly established. Moreover, the 'diplomatic' flavour is presented as .pdf files incorporating characters from multiple fonts, which impedes portability between word processors. By contrast, while the programme has opted not to use an international standard like Extensible Markup Language (XML), the 'base' flavour is encoded in .txt files using only the ASCII character set from which other flavours and formats can be generated; this flavour originates with *A Linguistic Atlas of Late Mediaeval English* (LALME) and is used by its other daughter projects too. The bulk of the MELD materials has never before been transcribed, let alone made public, and it is no small achievement of the MEST staff that they have visited a very considerable number of English local archives over several years in order to identify and photograph relevant documents for subsequent transcription.

LALME's editors discarded documents on account of their tending to be formulaic, lexically poor, or not evidencing local or regional usage. Moreover, LALME's localisation procedure fitted texts into a presumed continuum. It did so relative to other texts based on their linguistic forms (except for 'anchor texts' whose position on the continuum reflects extra-linguistic criteria) and disregarded any textual and historical factors that could explain the observed linguistic variation (p. 4). Last, the atlas maps are anachronistic since many of the texts date from the very end of the period the questionnaire was designed to cover (1350–1450) or even beyond it. By contrast, as discussed by Stenroos and Thengs in chapter 4, MELD includes only texts which explicitly state their localisation, although even such statements vary in their reliability; MELD operates with three levels ('explicit', 'historical' and 'inferred').

https://doi.org/10.1017/S1360674322000260 Published online by Cambridge University Press

The MEST team has developed robust criteria for the classification and subclassification of documents into functional categories. It is a document's function which selects its physical and linguistic forms as well as its language. Latin remains the default language selection for monolingual documents into the 1500s with a steady increase in the proportion of documents issued in monolingual English over time. Latin formulae embedded in an English matrix are common while the opposite composition is exceedingly rare, and a local document may at once mix regressive and progressive English forms. Impressionistically, the amount of English varies by function: documents more likely to contain English are the least formulaic ones and those addressed to lay audiences.

It is in clerks' development of fixed English forms of formulae – orthographically, morphologically and lexically – that English gradually standardises, argues Stenroos in chapter 5, with them transferring the conventions known to them from their Latin training on to English documentary texts. This transfer explains why Latin-trained clerks use digraphs ending in <-h> such as <wh>, <th>, <gh>, <ch> in place of the distinctly English graphs <ƿ>, <þ> and <ȝ> – the English graphs were, put simply, never a part of their active repertoire (pp. 126ff.). Since most scribes copying English-language literature did not share clerks' Latin training, a stark linguistic contrast emerged between literary and documentary texts as a major characteristic of the late and post-medieval English periods (p. 101). Standardisation had, accordingly, no single geographical locus. Moreover, LALME's editors described southern documents as having been standardised by the mid fifteenth century, that is to say void of locally or regionally marked forms but not necessarily invariant. Stenroos proceeds to argue that this description does not hold entirely true of MELD texts. This chapter is the first to present empirical linguistic data in any noteworthy quantity. It opens the first of the volume's two applied parts, part II, titled 'Text communities and geographical variation'.

Chapters 6 and 7 illustrate the first of two dimensions to the MEST programme's philosophy mentioned above. The former chapter, by Bergstrøm, compares the East Midlands part of the MELD corpus with a body of documents associated with Cambridge and detects various contrasts in the relative frequency of forms manifesting linguistic variables between them. For example, the forms <qw>, <w> and <wh> for the variable (wh), and <any>, <eny> and <ony> for ANY occur in proportions that set Cambridge, as a text community, apart from the rest of the East Midlands. Some forms belong further north, which suggests a presence of northern students at the city's university. Other forms are more typically found further south, which may indicate that supralocalisation processes reached Cambridge earlier than the surrounding countryside through urban hopping. The latter chapter, by Thengs, also addresses contrasts in the relative frequency of forms. The contrasts are between two Cheshire towns, Nantwich and Knutsford. The North-East Cheshire part of the MELD corpus provides a baseline, and the variables include many of the same ones addressed in other chapters, for example (th), (gh) and CALL VS CLEPE. Supralocal forms feature more prominently in the first of these towns. Their presence there tallies well with it being a

https://doi.org/10.1017/S1360674322000260 Published online by Cambridge University Press

market town situated on a major trade route, compared to the other town being more geographically isolated and, accordingly, less amenable to linguistic innovation.

A subcorpus of MELD 2016.1 amounting to 141 documents (~101,537 words) has as its principal function to define and describe land holdings, where many word-geographical studies have not similarly controlled for the factor of genre. Chapter 8, by Stenroos, tentatively identifies correlations between geography and the lexis used to define and describe land holdings and their geographical extent, including units of measurement. The reasons for the author's tentativeness towards the salience of the factor include low numbers of occurrences and difficulty establishing synonymy; thus the figure 8.4 map suggests *flat* is northern and *pightle* southern but it is based on a mere 4 occurrences of the former and 2 of the latter. Similarly, the numbers of occurrences respectively considered for figures 8.3, 8.5, 8.6 and 8.7 are 14, 8, 11 and 12.

Part III removes the focus away from possible correlations of variation in lexis and spelling with contextual factors like text communities, scribal communities and sponsors of literacy, which logically exhibit degrees of collinearity. This part convinces readers that documents are structurally and linguistically formulaic and suggests that social and pragmatic factors are what select code-switches between Latin and English and deployment of punctuation. It comprises three chapters, the first of which, chapter 9, by Jeremy Smith, takes advantage of how the MELD transcriptions stand out from other corpora by having recorded punctuation marks: qualitative analysis of selected documents shows them not only to contain marks in the first place, contrary to popular belief, but also to have deployed them non-randomly as 'an aid to oral delivery' (p. 218), albeit each in its own way. For example, D0124, a lease, has as its sole mark 'a punctus placed before the identification of the key issue, *viz.* the rent to be paid' (p. 211), while punctus and *litterae notabiliores* have various uses in D0167, among which is marking a name or other key word, the beginning of a rhetorical unit, or the end of an opening formula of address.

Abjurations record, in principle, statements given orally and in English by individuals being tried for heresy (Lollardy). In reality, however, they are not true records of words spoken, let alone those of the accused, possibly because they were taken down by religious authorities for them to pass on to secular authorities for sentencing. Comparison reveals them to comprise similar contents, structurally and linguistically. Chapter 10, by Kenneth Harestad-Solheim, divvies up a gradient into categories ranging from the fully formulaic with a fixed wording, at the one pole, to loose adherence to a template and greater variation in the exact wording, at the opposite pole. The basis, theoretically informed by Alison Wray's (2008) Morpheme Equivalent Unit as a yardstick for linguistic formulaicity, is 30 abjurations dated 1457–1509 and clustering in specific bishop's registers and regions. This clustering prompts the author pertinently to discuss whether the categorisation can be generalised to other abjurations. The chapter nonetheless provides a valuable contribution through its further underpinning one of the volume's principal takeaways – the formulaic language of documents – and through its constituting a first foray into uncharted territory: abjurations constitute, as far as this reviewer is aware, a body of materials that has

https://doi.org/10.1017/S1360674322000260 Published online by Cambridge University Press

received little attention from linguists, except to the extent that pragmaticists and semanticians have studied the superordinate that is confessions.

The volume's final chapter convincingly shows that 'mixed-language texts were common and an acceptable form of written language' (p. 267), with what its authors refer to as 'multilingual events' being non-randomly distributed and having specific functions. The authors, Stenroos and Delia Schipor, differ from previous scholars in anchoring their conception of such events, 'instances of written language mixing' (p. 254), in literacy studies rather than exclusively in linguistics and in concentrating on the pragmatics of language selection rather than the morphological or syntactical embedding of code-switches. Many document classes follow a template, with identifiable sections presented in a particular order and containing formulaic wording. Multilingual events, the chapter shows, most commonly relate to such formulae and occur in practically all document classes, although they are unevenly distributed among them, with powers of attorney containing many and letters few. The events flag textual elements such as sections, often in a visual and predictable manner, and their exact form exhibits greater variation when they are written in English than in Latin. Moreover, the High status of Latin may explain why these events are far more frequently encountered in documents whose matrix language is English than in those whose matrix language is Latin. It is the MELD corpus and documents housed in the Hampshire Record Office that the chapter mines for examples. The chapter thus demonstrates the presence of multilingual events outside 'macaronic' business texts produced at London, which previous scholarship has sometimes considered an anomalous group in the linguistic landscape of late medieval England.

A bibliography, a list of cited documents (giving their production date and localisation but not their functional class) and an index conclude the volume.

Of the volume's several virtues, an important one is how it relies on hard-and-fast criteria for document classification developed by the MEST team (chapter 3). Previous efforts at classification have certainly been both more *ad hoc*, impressionistic and less comprehensive. Another virtue is how the volume draws attention to documents' formulaicity in terms of both their language and their structure. Yet another virtue is that the volume charts new territory while being in line with current trends through its orientation towards sociolinguistics and pragmatics: for the contributors, the set of factors that might explain linguistic variation present in historical English includes such contextual ones as text communities and sponsors of literacy; and they accept multilingualism as being a fundamental characteristic of the late medieval period as far as both language users and texts are concerned. Multilingual texts should, accordingly, be embraced rather than dismissed. A principal finding is that documents attest several linguistic changes earlier than do literary texts, with the added nuance that the type and amount of multilingual events vary by document class: Latin dominates for a highly formulaic class like a power of attorney, while English is the matrix language for petitions and complaints. It makes good sense to propose as a route to standardisation the transfer of Latin conventions on to English.

https://doi.org/10.1017/S1360674322000260 Published online by Cambridge University Press

The most notable shortcoming is a disconnect between theory and linguistic data, especially when it comes to interpretation of quantitative such data. A set of empirical observations will contain clusters, and there will be correlations between such clusters and factors that will appear significant but in fact are random, or a factor may indeed explain some of the observations but another factor may be of greater intrinsic interest through it explaining more of the observations. Low absolute numbers of observations and/or high numbers of factors will increase the number of such apparent correlations. Sociolinguists have been subjecting their data to multivariate statistical testing since the 1970s to determine whether correlations are random and if they are not, what percentage of the observations a factor explains (its effect size) and whether more than one factor explains the same observations (collinearity). Within historical dialectology, already Michael Benskin (1988, 1994) called attention to the promise of digital corpora for collecting quantitative data and robustly testing them by quantitative means. The volume under review is by no means up to speed in this regard. It occasionally presents distributions of linguistic forms on geographical maps but what it more typically presents is their relative occurrence by factor in bar charts, sometimes stacked ones, at other times not. These graphic representations each relate observations to a single factor only, and the authors employ visual analysis of them to pronounce that factor salient. They present circumstantial evidence to dismiss other possible factors or do not discuss them at all. Such impressionistic analysis, which especially characterises part II (chapters 5–8), must accordingly be treated with a healthy measure of scepticism.

To take a concrete example, figure 7.11 gives the percentage occurrence of <th>, <y> and <þ> as realisations of word-initial (th) in all words except various function words. The figure contrasts their occurrence in Knutsford with their occurrence in North-East Cheshire, and what it reports separately for each location is a mean for several texts. The respective means are similar at 78/13/9 vs 93/1/6. It would have bolstered the argument for the author to have considered the variance for each of the two data sets – that is to say, for him to have taken into account how far each text is from the mean for its location – to help assess whether Knutsford in actual fact stands out from the North-East Cheshire baseline or the difference in means is merely a sampling effect. Another example is the percentage occurrences 36/44/21 vs 44/34/22 for <any>, <eny> and <ony> in table 6.3, which forms the basis for an argument that Cambridge is ahead of the East Midlands as a whole in its use of <eny>; however, a t-test (performed on the absolute occurrences) shows the difference not to be significant.

In fairness, Stenroos hesitates about the land holdings data as noted above. Moreover, the total numbers of texts mined for linguistic data have been added to several of the bar charts and a handful of figures and tables are accompanied by a footnote reporting the outcome of simple, univariate statistical tests like χ^2. However, they typically merely state the p-value without any further commentary in the footnote itself, let alone the body of the text. Given, in addition, that footnotes are sparse throughout the volume and only some distributions have been so tested, the impression is that the testing has been carried out as an afterthought, presumably at reviewers' request. The volume

https://doi.org/10.1017/S1360674322000260 Published online by Cambridge University Press

would, in short, have benefited from the chapters demonstrating the proposed factors' salience, effect size and non-collinearity more solidly.

Juan Camilo Conde-Silvestre, in his review of the volume, finds the conclusions 'sometimes obvious and not unexpected, clearly derived from the characteristics of the material under scrutiny, which occasionally may be a source of circularity' (2021: n.p.). This reviewer agrees. For example, it is hardly surprising that a body of materials excluded from LALME should show differences from the body of materials considered for LALME, nor that the differences should revolve around how supralocal forms are. And, if it is a document's function which selects its physical and linguistic forms as well as its language, it is at once by its physical form and its formulae that researchers recognise a document's function. Furthermore, Conde-Silvestre implies an absence of cross-fertilisation through the volume exclusively comprising chapters written by members of a close-knit research group and revisiting the same variables in several of them. Juan Manuel Hernández-Campoy and Conde-Silvestre (2015) (Paston Letters) and Moragh Gordon (2017) (Bristol civic documents and letters) both study variant realisations of one of these variables, the (th) variable, while Jacob Thaisen (2019) (MEG-C) discusses the <y> realisation of (th) as an example of variation falling in the interface between paleography and linguistics. It is in line with Conde-Silvestre's point about endogamy to note that the volume references neither of these recent studies. It would have been all the more pertinent for it to have done so since these studies also combine to put forward text-type, different types of literacy, training in Latin spelling conventions and the like 'non-conventional' factors as factors selecting realisations of the (th) variable, as opposed to geography.

The volume contains hardly any misprints, except 'Nantwich' is spelled with a lowercase *n* in the chapter 7 running head. Small infelicities are that the colours do not match between the data matrix and stacked bar chart in figure 6.9, and that what are described as green circles in the key to figure 8.4 are in fact blue.

Reviewer's address:
University of Oslo
P.O. Box 1003
Blindern
0315 Oslo
Norway
jacob.thaisen@ilos.uio.no

References

Benskin, Michael. 1988. The numerical classification of languages, and dialect maps for the past. In Pieter van Reenen & Karen van Reenen-Stein (eds.), *Distributions spatiales et temporelles, constellations des manuscrits*, 13–38. Amsterdam: John Benjamins.

Benskin, Michael. 1994. Descriptions of dialect and areal distributions. In Margaret Laing & Keith Williamson (eds.), *Speaking in our tongues: Medieval dialectology and related disciplines*, 169–87. Woodbridge: D. S. Brewer.

https://doi.org/10.1017/S1360674322000260 Published online by Cambridge University Press

Conde-Silvestre, Juan Camilo. 2021. Review of Merja Stenroos & Kjetil V. Thengs (eds.), *Records of real people: Linguistic variation in Middle English local documents*. https://linguistlist.org/issues/32.3987 (accessed 12 July 2022).
Gordon, Moragh. 2017. The urban vernacular of late medieval and Renaissance Bristol. PhD dissertation, Utrecht University.
Hernández-Campoy, Juan Manuel & Juan Camilo Conde-Silvestre. 2015. Assessing variability and change in early English letters. In Anita Auer, Daniel Schreier & Richard J. Watts (eds.), *Letter writing and language change*, 15–34. Cambridge: Cambridge University Press.
McIntosh, Angus, Michael Samuels & Michael Benskin (eds.). 1986. *A Linguistic Atlas of Late Mediaeval English*. Aberdeen: Aberdeen University Press.
Thaisen, Jacob. 2019. The round allograph of <r> in late Middle English. *Studia Anglica Posnaniensia* 53, 129–44.
Wray, Alison. 2008. *Formulaic language: Pushing the boundaries*. Oxford: Oxford University Press.

(Received 14 July 2022)

doi:10.1017/S1360674322000272

Erik Smitterberg, *Syntactic change in Late Modern English: Studies on colloquialization and densification* (Studies in English Language). Cambridge: Cambridge University Press, 2021. Pp. xii + 302. ISBN 9781108564984.

Reviewed by Christian Mair, Universität Freiburg

In his new book Erik Smitterberg investigates diachronic changes in four syntactic variables in the period from 1700 to 1900: *not*-contraction; co-ordination by *and*; nouns serving as premodifiers of other nouns; and participles postmodifying nouns. The first two variables were chosen as syntactic symptoms of colloquialisation, the last two as indexes of information compression (or, to use the author's preferred term: densification). The primary sources of data are the *Corpus of Nineteenth Century English* (CONCE) and the *Corpus of Nineteenth-Century Newspaper English* (CNNE). Comprising *c.* 1.3 million words in total, both are carefully compiled small corpora (by present standards of size). The corpus analyses presented in the book are traditional in the best sense of the word. They combine carefully compiled descriptive statistical surveys of corpus frequencies (with some additional multifactorial analysis) and philologically competent qualitative analysis of selected individual examples in their textual and historical context.

There is by now a large body of corpus-based research on colloquialisation and densification, and also on the specific variables in focus here. In such a situation, it is not the study design itself – solid and well thought out as it may be – that will provide the source of innovation. But there is ambition and an innovative thrust at a higher level, because the presentation of the corpus findings is complemented with much theoretical discussion that addresses some of the fundamental theoretical issues in

https://doi.org/10.1017/S1360674322000260 Published online by Cambridge University Press