Zeitschrift: SPELL : Swiss papers in English language and literature
Herausgeber: Swiss Association of University Teachers of English
Band: 44 (2024)

Artikel: Creating a corpus of late modern english pauper letters : uncertainties, challenges, and solutions
Autor: Auer, Anita / Gardner, Anne-Christine / Iten, Mark
DOI: https://doi.org/10.5169/seals-1053570

Anita Auer, Anne-Christine Gardner, Mark iten
(University of Lausanne)

Creating a Corpus of Late Modern English Pauper Letters: Uncertainties, Challenges, and Solutions[1]

The creation of searchable corpora and digital databases based on historical data can be challenging for various reasons, e.g. the lack of meta-data including the origin of a text or information about the writer, illegibility of the handwriting, and variant spelling. Within this context, this article discusses the challenges and uncertainties met, solutions found, and decisions taken for the creation of a corpus of pauper letters that were written under the Old Poor Law during the period 1795 to 1834. As the labouring poor received variable education before the introduction of compulsory elementary schooling in 1880, the pauper letters may have been written by the petitioners themselves or somebody else. This article therefore considers uncertainties related to the determination of sociolinguistic metadata, and as a result, the interpretation of the findings from a historical sociolinguistic perspective, as well as challenges linked to the dates of the letters, geographical anchor points, and the orthographic normalisation of the data. Our discussion reveals that despite the existence of other Late Modern English letter corpora and a good amount of existing expertise in the field, every new text type, and letter type in this case, comes with new challenges.

Keywords: corpus linguistics; corpus creation; English historical linguistics; Old Poor Law; pauper letters

[1] This article was written in the context of the SNSF-funded research project *The Language of the Labouring Poor in Late Modern England* (2020-2025; 100015_188879). Many thanks to Kilian Schindler, Julia Straub, and the anonymous reviewer for their valuable feedback on an earlier version of this paper. All remaining shortcomings lie solely with us.

In the field of English Historical Linguistics, the emergence and rapid increase of dictionaries and text corpora (online or in electronic form) as data sources for linguistic research can be observed from the 1990s onwards (see Bergs & Brinton 2012: xi; Kytö 2012; Stratton 2020). Since the compilation of *The Helsinki Corpus of English Texts: Diachronic and Dialectal* (Rissanen et al. 1991), which was the first electronic corpus in the field, many multi-genre and single-genre corpora covering different time periods have been created. While some of these corpora were projects in their own right, more specialised corpora have also been created to answer specific research questions as part of individual projects. Even though a good amount of corpus experience has been available in the research field for some time and there have been attempts to create standards, the individual needs of a researcher from a corpus, text type-specific characteristics, and the fast development of encoding standards and tools are only some aspects that can be challenging. Other challenges linked to the creation of searchable corpora and digital databases based on historical data are the lack of meta-data including the origin of a text or information about the writer, hardly legible handwriting, and variant spelling.[2] Couched within this context, this article discusses the challenges and uncertainties met, solutions found, and decisions taken for the creation of a corpus of pauper letters that were written under the Old Poor Law and cover the period 1795 to 1834. As the labouring poor received variable education (if any at all) before the introduction of compulsory elementary schooling in 1880, the pauper letters may have been written by the petitioners themselves or by somebody else, typically from their social circle, who was able to write (see Sokoll 2001; King 2019). Selected other challenges that we encountered concern dates of the letters, the determination/verification of a pauper's geographical origin based on phonetic spelling, and the orthographic normalisation of the data. Based on a range of illustrations concerning the latter points, we show that despite the existence of other Late Modern English letter corpora and a good amount of existing expertise in the field, every new text type, and letter type in this case, comes with new challenges. In line with this, we have found that a detailed description of the corpus creation process and the data-related challenges, which are often outlined in corpus manuals are of great importance and relevance for historical linguists as the respective decisions on for instance spelling or punctuation can have an effect on the interpretation of the data and therefore the respective research findings (see for

[2] The term 'variant spelling' is discussed in more detail in Gardner's (2023b) "Speech Reflections in Late Modern English Pauper Petitions."

instance Kytö et al. 2011 on ETED and the information provided on the website of *The Mary Hamilton Papers (1743-1826)*).

The article is structured as follows: section 1 provides a brief overview of the development of corpus linguistics in the field of English historical (socio)linguistics. Section 2 focuses on the corpus of the Language of the Labouring Poor in Late Modern England (LALP) and describes its history as well as the types of letters included. Section 3 is dedicated to uncertainties, challenges and solutions in the LALP corpus. Finally, Section 4 provides concluding remarks including thoughts on how to navigate uncertainties in the digital humanities.

1. Corpus Creation in the Field of English Historical (Socio)Linguistics

As previously mentioned, from the 1990s onwards, the field of English historical linguistics has seen a rapid development in the creation of multi-genre and single-genre corpora. These corpora vary in terms of (1) source material, notably manuscript or printed material, found in archives or based on existing editions, and (2) sampling, where a distinction is made between the use of (a) text samples (restricted number of words) representative of a certain variety, usually linked to textual or sociolinguistic criteria, and (b) convenience or opportunistic samples, i.e. based on data available to the researcher but lacking rigorous sampling criteria (see Nelson 2010: 57; see also Stratton 2020). Depending on the data to be compiled in a corpus, text type and social information or other potentially relevant meta-linguistic information is available. For instance, the field of historical sociolinguistics, which aims at gaining a better understanding of the relationship between language and society in the past by applying synchronic sociolinguistic theories combined with philological approaches, has led to a focus on ego-documents, i.e. sources such as autobiographies, diaries or letters (see Auer et al. 2015; Auer & Hickey in press). These text types can provide scholars with additional, e.g. social, information such as the age, gender, social class, education, occupation of the writer. This type of social information combined with linguistic factors can often explain language variation and change (see Auer et al. 2015). We want to illustrate this with a study of the linguistic variants *you were / you was* in Late Modern English data (Auer 2014). As the use of *you was* was considered "an enormous Solecism" by the grammarian Robert Lowth (1762: 48), the question may be raised whether grammatic-

al judgements of this type affected language use across the social stratum. In fact, in Auer (2014), we can see that the use of the standard form *you were* is strongly associated with well-schooled writers, while the *you was* variant is clearly dominant in lower-class writing. This finding suggests that the form emerged from below and spread from there into other social spheres. It is thus an important aim of historical sociolinguists to work with data that allow for the investigation of linguistic variation and change across the social spectrum.

As literacy was socially stratified in England, literacy rates in different time periods determined the availability of texts produced by different social groups (see Auer & Hickey in press). Correspondence data from the Late Middle English and Early Modern English periods that have served as linguistic data are for instance the Paston Letter Collection (see Bergs 2005; Hernández-Campoy & García-Vidal 2018; Hernández-Campoy 2021) and, most notably, the *Corpus of Early English Correspondence* (CEEC; Nevalainen & Raumolin-Brunberg 2017). For the Late Modern English period (c. 1700-1900), a range of letter collections and corpora are available to English historical sociolinguists, e.g. the *Corpus of Early English Correspondence Extension*, the *Corpus of Late 18c Prose*, *The Bluestocking Corpus*, and *The Mary Hamilton Papers*. As most of these corpora contain correspondence from the middle and upper layers of society (including people who were well known), a good amount of meta-linguistic, including social information has been readily available to the corpus compilers. Attaining this type of information becomes more challenging in relation to data produced on the lower end of the social stratum, as the following sections demonstrate.

2. The LALP Corpus

The data discussed in this article are pauper letters that were written and sent within the context of the Old Poor Law and that specifically cover the period 1795-1834. These letters have been converted into a searchable corpus as part of the SNSF-funded research project *The Language of the Labouring Poor in Late Modern England* (2020-2025), and they serve as the basis for investigating the role of social stratification in language variation and change during the previously mentioned period. Many of the pauper letters included in the corpus were originally collected by Tony Fairman from archives all over England (see Auer & Fairman 2013; Auer et al. 2014). Within the context of the project, facsimiles from the

archives served as the basis for diplomatic (re-)transcriptions of the letters from which meta-linguistic information such as type of letter, authenticity, sender details and role, writer details and role, recipient details and role, content function and objective was extracted. In addition to the diplomatic transcriptions, plain text, normalised, and xml versions of the corpus are currently being created. All the different stages outlined here require around four people (preparation, checking, double-checking), which may already indicate that the process is very time-consuming including regular discussions amongst the project members and additional historical research. Almost all of the letters have some challenges, notably at least the question of authenticity, i.e., whether the letter is autographical or written by somebody else.

As regards the specific text type, as previously noted, the letters were written in the context of the Old Poor Law (also known as the Poor Relief Act of 1601 or the Elizabethan Poor Law) that aimed at providing a system for poor relief distribution in England and Wales. Within this system, the parish was the central administrative office that had appointed overseers who were responsible for poor law legislation. During the period 1795-1834, the payment and receipt of out-relief from parish funds was legalised. This means that impoverished labourers, artisans, and other people who had lost their belongings and were 'in distress,' had the right to apply for out-relief to their parish of legal settlement by sending a letter, and relief was then offered in the form of money being sent or by removing the paupers from their domicile at the time and bringing them back to their parish of legal settlement, where they were typically placed in a workhouse (see Whyte 2004: 280; Auer & Fairman 2013). The text type associated with the Old Poor Law are thus application letters for poor relief that had to be written by the labouring poor who had often only received limited schooling.[3] We provide a facsimile and a (plain text) transcription of a pauper letter by Moses Tyson (10 September 1828) from Lancashire below for illustrative purposes:

[3] On the text-typological distinction between pauper letters and petitions see Gardner (2023c).

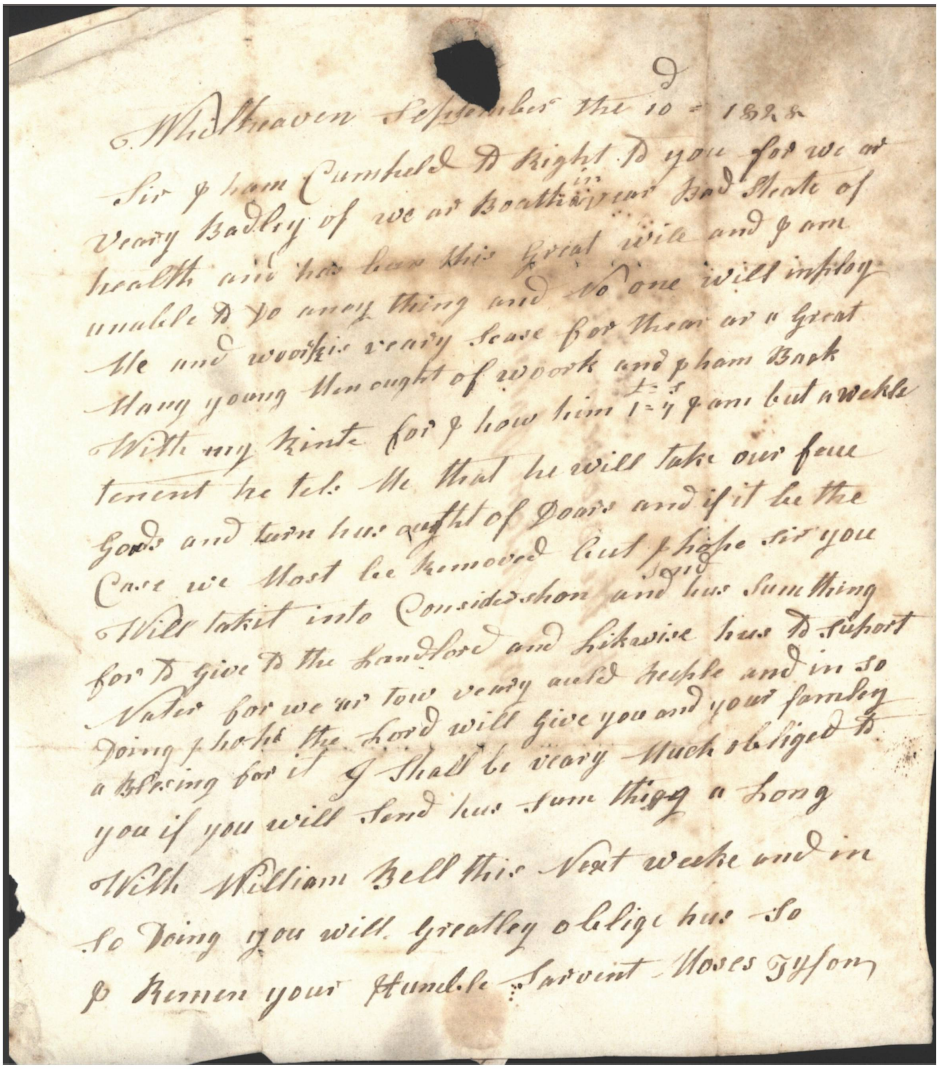

Figure 1. Facsimile of a petition letter, reused with permission of the Cumbria Archive Centre, Barrow-in-Furness (Ref: BPR10O52)

Whitheaven September the 10=d= 1828
Sir I ham Cumpeld to Right to you for we ar
Veary Badley of we ar Boath =in=[^INSERTION^] a vear Bad Steate of
health and has been this Great wile and I am
unable to Do aney thing and No one will inploy
Me and woork is veary Scase for thear ar a Great
Many young Men ought of woork and I ham Back
With my Rinte for I how him 1=£= = 7=s= I am but a wekle
tenent he tels Me that he will take our feue
Goods and turn hus ou[^g OVERWRITES t^]ht of Doars and if it be the
Case we Most be Removed but I hope Sir you
Will takit into Considershon and =Send=[^INSERTION^] hus Sumthing
for to Give to the Landlord and Likwise hus to Suport
Nater for we ar tow veary ould people and in so

> Doing I hope the Lord will Give you and your famley
> a Blesing for it I Shall be veary Much obliged to
> you if you will Send hus sum thi[^n OVERWRITES {g}^]g a Long
> With William Bell this Next weeke and in
> so doing you will Greatley oblige hus So
> I Remen your Humble Sarvent Moses Tyſon

In comparison to letters written by well-educated members of the society, the example of the pauper letter above suggests that the writer of the poor relief application has not received any grammatical schooling. The example above does not, for instance, contain any punctuation and the capitalisation of words is random. Moreover, the spelling varies (*ham* and *am*), often containing reflections of spoken language (e.g. *hus* for *us*, *sum thing* for *something*, *Rinte* for *Rent*), often not adhering to forms codified in contemporary dictionaries and spelling books. Similarly, as the focus of lower-class writing training was on mechanical aspects in the first instance, grammatical rules were likely not taught, but they may have been picked up on through other available sources like the Book of Common Prayer and the Bible. This is in fact one of the questions that *The Language of the Labouring Poor in Late Modern England* research project aims to answer.

The corpus does not only contain poor relief application letters and thus official correspondence but also letters written by paupers to family members or other people within the same social sphere (i.e. the *Pauper Letter Core Corpus*); in addition, a related correspondence sub-corpus contains correspondence between overseers and letters written by other people, e.g. landlords, neighbours, to support or denounce an applicant for out-relief. The main corpus that allows us to shed new light on the language of the labouring poor – the *Pauper Letter Core Corpus* (PLCC) – consists of approximately 820 letters (c. 150'000 orthographical units)[4] from all English counties whereby the distribution varies: this is dependent on the availability of data. For instance, the corpus contains only 1 letter from Lincolnshire, but it contains 109 letters from Essex. While the greatest part of the corpus consists of one or two letters sent by a pauper, the corpus also contains material by paupers who have sent several letters, e.g. Frances Soundy (20 letters, Berkshire), Charls Ann Green (8 letters, Dorset), and Augustine Morgan (8 letters, Dorset). Most of the letters addressed to overseers petition for financial support, i.e. they are thus

4 As we are still working on the corpus, and decisions of categorisation are still being reviewed, these numbers may change slightly.

addressed to someone higher up the social scale. There is only one case in the corpus so far where the pauper, notably Margret Lee from Durham, has sent two letters of which one is addressed to an overseer and the other to her father; this type of data gives us the unique opportunity to compare the language use of the pauper when addressing people from different social layers, discussing different topics, and to shed light on the pauper's written repertoire.[5]

3. LALP Corpus: Uncertainties, Challenges, and Solutions

The nature of the data has led to several challenges during the corpus creation process. Many of these concern uncertainties related to the writers of the letters and the effect that this has on the determination of sociolinguistic metadata, and as a result, on the interpretation of the findings from a historical sociolinguistic perspective. Other challenges are linked to the dates of the letters, geographical anchor points, and the orthographic normalisation of the data. In the following sections, we present several of these challenges along with solutions on how to overcome them.

3.1. Authenticity

One of the main uncertainties is related to the authenticity of the pauper letters. It is difficult to ascertain who physically penned a letter, as the writer does not always correspond to the sender of the letter, which we defined as the person whose name appears in the signature of the letter. Since many of the labouring poor generally had only limited exposure to elementary education, we cannot assume that every applicant had the ability to write to their home parishes themselves.[6] In some rare cases, the letters themselves hold information concerning the writing abilities of the sender, e.g. "I […] had no Lerning but what my poor Mother Gave me to Read or wright" (DOR_1822_001). Further archival work which would allow us to compare the handwriting of the letters with that documented in parish registers, if extant, unfortunately lay outside the scope of the project. Therefore, we only considered letters as autographical when they

5 For more detailed linguistic analyses of pauper letters see Laitinen and Auer (2014), Auer et al. (2023), Gardner (2023b; 2023c), and Gardner et al. (2022).

6 For details on schooling opportunities for the labouring poor see e.g. Auer et al. (2023) and Gardner (2023a, 2023c).

contained conclusive metalinguistic comments, e.g. regarding the writing process or their education, as shown in the earlier example.

However, even in the absence of such concrete evidence most of the pauper letters sampled for the corpus can still be regarded as (likely) authentic, i.e. representative of lower-class writing. In their extensive studies on large numbers of pauper letters, which also in part permitted verification in parish registers, historians Thomas Sokoll (2001: 64–65) and Steven King (2019: 36–37) conclude that most of the letters were written by the applicants themselves or someone from their social circle with limited schooling. This is particularly probable when several letters by the same sender are written by the same hand and the correspondence covers a longer period of time.

In order to classify letters as likely authentic or non-authentic we introduced the non-linguistic criterion of level of handwriting. When the characters, lines and ink flow are rather uneven (as in the facsimile above), we assume the writer to have comparatively little training and therefore that the letter is likely authentic and representative of lower-class writing. In contrast, when the handwriting and layout is very regular, perhaps even involving flourishes, we argue that the writer was highly trained and the letter is consequently regarded as non-authentic (for images contrasting these two types of handwriting, see Sokoll 2001). Anne-Christine Gardner (2023a) has shown that the non-linguistic differentiation between (likely) authentic and non-authentic pauper letters is also reflected on a linguistic level. For example, in authentic letters penned by less-trained hands we find a wider range of closing formulae (e.g. based on *I remain*, *from* or *no more*) and self-references (e.g. *yours to command* besides the common *your humble/obedient servant*), some of which were specifically associated with the lower classes. On the other hand, closing formulae (e.g. *I am*) and self-references (e.g. involving *servant* or missing altogether) in non-authentic letters, i.e. those penned by well-schooled individuals, correspond more closely to usage patterns found in business correspondence at the time. The distinction between (likely) authentic and non-authentic pauper letters is particularly useful when several missives survive by the same sender, but at least two different hands were involved across multiple letters. In this case we assume that none of the letters are autographical, but determine for each letter on the basis of the handwriting whether it is likely to have been authentic or not. Very rarely, only a copy of a pauper letter survives, for instance because it was copied by an overseer to keep in their records. Here we cannot say with any certainty

whether the letter is authentic since we cannot be certain that it was copied faithfully, particularly with respect to the spelling.

As the distinction between authentic and non-authentic letters, and thus between sender and encoder, is linguistically meaningful, we provide metadata entries not only for the sender, but also for any encoders involved. However, determining relevant sociolinguistic metadata of the possible encoder(s) is challenging. When references within the text indicate that a family member penned the letter, rather than the signatory, we may have some limited data to include in the metadata entry; in these cases, we often rely on additional information provided in other letters by the same sender or household, if available. When it is clear that someone else outside the family circle wrote the letter on behalf of the sender, there is often no sociolinguistic metadata concerning the encoder available at all, especially when letters are classified as non-authentic. Challenges and solutions surrounding the authorship and authenticity of pauper letters are discussed further in Gardner et al. (2022) and Gardner (2023a; 2023c).

3.2. Dates

A second challenge regarding the creation of the metadata entries for the letters relates to the date on which the letters were written. For the project, we only include letters written between 1795 and 1834 in our *Pauper Letter Core Corpus*, which means that it is important to know the individual letter dates – or at least the year – so that we do not include older or newer letters by accident. The majority of the letters contains the date within the text of the letter itself, usually either as part of the header or the footer, often together with the location or current domicile of the sender. Sometimes, the pauper does not write the date in that manner, or only includes parts of the date, e.g. only day and month, only the year, only the month and year, etc. Due to damage, overwriting, unclear writing, and the like, parts of the date may also not be identifiable by the project team, and in the case of overwriting, we do not always know which character was the one doing the overwriting as opposed to being overwritten. Additionally, in very rare cases, more than four digits are found in the header of the letter, e.g. year noted as *18011*, so that it is not clear whether they meant *1801* or *1811*. In this particular case, the project team decided to keep the decade slot free, thus 18X1. As the date information is also reflected in the FILENAME, which is made up of the county abbreviation,

the year, and the letter number, in this specific case, the filename is DOR_18X1_001.

In some cases where the date of the letter was not specifically mentioned by the sender themself or if the date is not certain due to other factors, the postmark is a helpful tool to determine the approximate date of writing, which usually ended up being the day of or in the days following the writing of the letter itself, as a stamp was used to show the date and location of the post office that the letter went through. Then again, the stamps and the dates written in the main text do not always match. A letter from Surrey (SUR_1800_002), for example, contains a stamp dated April 13, 1801. However, the author of the text writes "Ap[^r OVERWRITES e^]eal sunday the 12 1800," which seems to be incorrect as proven by the stamp, as well as by the fact that April 12 was a Sunday in 1801 and not in 1800. Whenever stamps or other external information are not available to us, we have to trust the accuracy of the letter writer to include the correct date.

Additionally, there are some options that help determine a certain range of dates in which the letter must have been written. The material the letter was written on sometimes contained a watermark including the year in which the material was produced. In those cases, we assume that the letter was, at the earliest, written in the year shown in the watermark. However, this does not help with the upper boundary of 1834.

At times, the content of the letters helped as a clue as well, e.g. when the letter references a certain holiday, such as Christmas, Easter, Michaelmas, or similar, or when it references a previous letter that itself does show a specific date, for example "my letter from two weeks ago." When the sender has sent multiple letters, it can sometimes be deduced where the letter falls chronologically among the rest of the, hopefully dated, letters. Very rarely, the date may be determined by outside sources such as information from the archive that could not be found in the manuscripts available to the project team.

In any case, whenever the date could not be easily determined, we made a note of how we decided on a date in the metadata information, e.g. information from postmark, from archive, etc. Overall, only a few letters remain completely undated, which were then removed from the *Pauper Letter Core Corpus*.

3.3. Geographical Anchor Points

In Section 3.1., we have already provided some information related to the challenges linked to the sender's and the writer's authenticity. While the main aim of the project is to shed light on the role of social stratification in language variation and change based on pauper letters and the social information is therefore important, the fact that application letters for out-relief were written all over England also raises the question whether the language of the letters reflects geographical differences. In most cases, we know where the letter was written from, if it was included in the header or footer of the letter, and we also know the apparent parish of legal settlement, i.e., where the letter was sent to. Challenges can be linked to the place from which the letter was sent, the parish it was sent to, and the geographical origin of the pauper, the latter of which may provide insight into the dialectal origin of the pauper as well. As regards the places from and to which letters were sent, the pauper's spelling of the name can make it difficult to determine the location. For instance, *Newinkless* in Birmingham, which was mentioned in a letter from 1829 in the Gloucestershire letters, turned out to be *New (H)In(c)kleys*, a street in Birmingham that disappeared in 1852 when the area was razed to the ground in order to build New Street Station for the London and North-Western Railway.[7] Another example concerns the sender's address of *Wesbe{a}ch* mentioned in a Northamptonshire letter in 1826, which we determined to be Wisbech as it was geographically closer to the parish of legal settlement than *West Beach*; moreover, this decision was then confirmed by the work of the social historian Steven King (2009). Similarly, in the Staffordshire letters, we came across *Etelay Street* that turned out to be *Heatley Street* in Preston, thus displaying an example of H-dropping. Moreover, the Robin Hood in Windmill Street, *heamarkett*, was identified as *Haymarket* in the City of Westminster, notably referring to the place of a pub that was demolished in the 1880s ("Robin Hood").[8]

As regards the parish of legal settlement, an affiliation was required for a pauper in order to be entitled to receive poor relief, according to seventeenth-century statutes (see Whyte 2004: 280). It can however not be assumed that the parish of legal settlement is necessarily the same as a

---

7 More information on this is available on the website of the British Museum ("New Inkleys"), found at www.britishmuseum.org/collection/term/x47943, accessed on 23 October 2023.

8 For information see www.pubology.co.uk/pubs/7467.html, accessed on 12 September 2023.

sender's geographical origin, which may shed some light on a person's dialect use, as the following extract by Whyte (2004) reveals:

> Settlement rights could be established on the basis of birth, marriage, and, in the nineteenth century, from a father's or even grandfather's parish of settlement. Other mechanisms, such as renting property worth £10 per annum, a year's agricultural service, completing an apprenticeship, paying taxes or serving in a parish office for a year were also grounds for gaining a settlement. People who required poor relief and were living in a parish which was not their parish of settlement could be removed there or, less commonly, be provided with out-relief. (280; see also Auer & Fairman 2013: 11)

Given the different ways of obtaining a legal settlement, notably either through marriage (as was always the case for women) or living at a location for a certain length of time, or similar circumstances, potential geographical and dialect origin need to be determined in a different way. A previous study of pauper letters sent to parishes of legal settlement in Dorset has shown that paupers migrated extensively, typically in search of work, and on average moved far greater distances than any other socioeconomic group at the time (Gardner et al. 2022). Only few of these paupers had remained in Dorset – most had migrated to neighbouring counties or London.[9]

In some cases, certain phonetic spellings can give us a hint as to whether the pauper's geographical origin can be compared to a region's dialectal features in their speech. For instance, a study by Gardner et al. (2022: 61–65) of seven letters written by the pauper Moses Tyson in Cumberland between September 1828 and February 1830 reveals that while the letters contain speech reflections like *dun* and *sumthing* for *done* and *something* (STRUT vowel), raising of /e/ to /i/ as in *prisent* for *present, our Rent and Coals is,* and H-insertion as in *hus* for *us* and *hever* for *ever*, some of which are indicative of a Northern origin, it is difficult to determine clear-cut regional dialect boundaries. Similarly, in a study based on 31 pauper letters sent to parishes in Dorset between 1742 and 1834, Gardner (2023b) identified 52 different phonological and morphosyntactic features which in fact link the writers to Dorset and/or the South West more generally. These features include KIT lowering and FLEECE shortening, which are suggested by spellings such as 'famely' for *family* and 'wick' for *week*, respectively. Linked to the geographical/dialect ori-

[9] Sokoll (2001: 32–43) made similar observations in his substantial investigation of Essex pauper letters.

gin challenge, it can generally be observed that the more grammatical schooling the writer received and the more experienced they are, the closer their language use is to the written standard, and the more difficult it is for us to determine their origin.

3.4. Orthographical Normalisation

Another challenge is linked to orthography, as already indicated in the previous section linked to street and place names. Considering that the letter writers likely did not receive much schooling, orthographic variation is often found in letters, sometimes to the point where a word could only be identified thanks to the content and context. As illustrated in the pauper letter in Figure 1 in Section 2, the diplomatic transcriptions and plain text versions of the pauper letters maintain original spelling and word boundaries. Given that spelling and word boundaries are often highly idiosyncratic and variable, this creates a problem for concordance tools and other search interfaces. Therefore, it is necessary to prepare a normalised version of the corpus, which can be done with VARD, a variation detector tool that detects, normalises and tags variant spellings (Baron & Rayson 2009). Here is an illustration of what the normalised form of the word *goon* for *gone* looks like (Auer et al. 2014: 20):

<normalised orig="goon" auto=false">gone</normalised>

As the tool was originally developed to deal with Early Modern English spelling, it could only detect a limited amount of variations in the pauper letters at first. VARD has however since been updated and "employs techniques from modern spell checking software to search for potential variants and find candidate equivalents for variants found" (Baron & Rayson 2009). In addition to being able to detect phonetic spelling more easily, the programme can also be trained to recognise certain variants and to supply corresponding variants. Having said that, manual checking is still necessary as some words are not recognised as variant spelling since the words correspond to modern standard variants, e.g. the words in bold in the examples below from Lancashire:

(a) Sir I **ham** veary Sorey
(b) my Dother Child pention whitch **his** 13 wekes
(c) if it be the Case we **Most** be Removed

(d) for She **as** been 5 weeks
(e) if you will Send hus **sum** thing

As indicated in (e) with *sum thing* above, another challenge concerns word boundaries: the letter writers often separated words that are joined together according to modern standard spelling. Other common examples are *a Gain* for *again* and *be hind* for *behind*, notably words that contain common elements like *a* and *be*, which the writer may interpret as the indefinite article and the auxiliary verb, respectively, and therefore does not join up. Similarly, in the latter two cases VARD detected all single elements, i.e. *a, gain, be*, and *hind*, as these words exist in its lexicon, but it did not recognise the unjoined elements as variant spelling of *again* and *behind*. Despite these spelling-related challenges and the fact that double- and triple-checking is required, VARD has sped up the normalisation.

In contrast to the diplomatic and the plain text versions of the corpus, word-internal overwritings and insertions are not marked, so only the final version of the words end up in the normalised version of the corpus. On the other hand, cancelled material such as crossed or inked out parts of the text is marked as *[-TEXT-]* as long as it concerns full separate morphemes.

Sometimes, certain common abbreviations, such as *Mr*. for *Mister* or *inst*. for *instant* appear so frequently within the *Pauper Letter Core Corpus* that we keep them in their abbreviated forms in our normalised version. In order not to lose track, a list containing all such abbreviations is in the works.

Even though we are able to normalise almost every orthographical unit in the corpus with the help of VARD and manual checking, some words remain unidentifiable or unclear. Whenever the project team is able to make an educated guess, unclear words are normalised as such and are marked with an asterisk. In cases where not even guesses are possible, the initial transcription is kept and marked with a paragraph sign. Sometimes, the project team decided to preserve the original transcription in brackets, for example, when a verb form was used that does not exist in Present-Day English, e.g. *seed* used as a form of *saw* will be normalised to *saw [%seed]*. Given that it is impossible to provide a great amount of additional information in the corpus files, the project team suggests that future users of the corpus consult the diplomatic transcription and, if possible, the original facsimiles for any of the aforementioned special cases. Moreover, a manual including detailed information on the transcription and meta-data choices will be made available with the corpus.

4. Concluding Remarks

It was the aim of the current article to describe and illustrate uncertainties and challenges linked to the creation of an electronic corpus based on letters written by paupers in the context of the Old Poor Law during the period 1795-1834. Even though a great amount of expertise related to the creation of historical electronic corpora in the field of English linguistics already exists – including expertise linked to philological aspects (orthographic variation, self-corrections and other corrections, capitalisation, punctuation, line breaks, etc.) – and we were able to rely on previous practices, such as David Denison (1994) on correspondence and Merja Kytö et al. (2011) on depositions, every new text type comes with new challenges for which solutions have to be found. Despite the existence of correspondence collections and corpora from the Late Modern English period, the focus on the lower orders who did not receive grammatical schooling has introduced numerous uncertainties for which the project team needed to find solutions, such as those linked to the authenticity of the letters (sender versus writer and respective biographical information) as well as dates. Orthographic variation in the letters as a result of incomplete (grammatical) writing training and experience did not only make the determination of place names difficult, but also created numerous challenges during the orthographic normalisation task of the data. The rapid development of artificial intelligence will likely give rise to new tools that can be used for the transcription and annotation of historical manuscript data. At the same time, many of the challenges and uncertainties we have faced with respect to the pauper letter data have been and could only be solved by relying on existing practices that were documented in great detail and many long discussions within the project team and with international colleagues. Going forward, we similarly document the challenges and decisions in much detail so that future linguistic corpus compilers and other scholars in the field of digital humanities have some more guidance in their data-related decision-making processes.

References

Auer, Anita. 2014. "Nineteenth-Century English: Norms and Usage." *Norms and Usage in Language History, 1600-1900. A Sociolinguistic and Comparative Perspective*, edited by Gijsbert Rutten et al., John Benjamins, pp. 151–170.

Auer, Anita, and Tony Fairman. 2013. "Letters of Artisans and the Labouring Poor (England, c. 1750-1835)." *New Methods in Historical Corpora*, edited by Paul Bennett et al., Narr Verlag, pp. 77–91.

Auer, Anita, et al. 2014. "An Electronic Corpus of *Letters of Artisans and the Labouring Poor* (England, c. 1750-1835): Compilation Principles and Coding Conventions." *Recent Advances in Corpus Linguistics: Developing and Exploiting Corpora*, edited by Lieven Vandelanotte et al., Rodopi, pp. 9–29.

Auer, Anita, et al. 2015. "Historical Sociolinguistics: The Field and Its Future." *Journal of Historical Sociolinguistics*, vol. 1, no. 1, pp. 1–12.

Auer, Anita, et al. 2023. "Patterns of Linguistic Variation in Late Modern English Pauper Petitions from Berkshire and Dorset." *Intra-Writer Variation in Historical Sociolinguistics*, edited by Markus Schiegg and Judith Huber, Lang, pp. 133–156.

Auer, Anita, and Raymond Hickey. In press. "Ego Documents in the History of English." *New Cambridge History of the English Language*. Vol. 2: *Documentation, Data Sources and Modelling*, edited by Merja Kytö and Erik Smitterberg, Cambridge UP.

Baron, Alistair, and Paul Rayson. 2009. "Automatic Standardization of Texts Containing Spelling Variation: How Much Training Data Do You Need?" *Proceedings of the Corpus Linguistics Conference* (CL2009), University of Liverpool, UK, 20–23 July 2009, edited by Michaela Mahlberg et al.

Bergs, Alexander. 2005. *Social Networks and Historical Sociolinguistics*. De Gruyter.

Bergs, Alexander, and Laurel J. Brinton. 2012. "Preface to the Handbook of English Historical Linguistics." *English Historical Linguistics: An International Handbook,* edited by Bergs and Brinton, vol. 2, De Gruyter Mouton, pp. xi–iv.

Corpus of Early English Correspondence (CEEC). 1998. Compiled by Terttu Nevalainen, Helena Raumolin-Brunberg, Jukka Keränen, Minna Nevala, Arja Nurmi and Minna Palander-Collin at the Department of Modern Languages, University of Helsinki.

Corpus of Early English Correspondence Extension (CEECE). 1998. Compiled by Terttu Nevalainen, Helena Raumolin-Brunberg, Samuli Kaislaniemi, Mikko Laitinen, Minna Nevala, Arja Nurmi, Minna Palander-Collin, Tanja Säily and Anni Sairio at the Department of Modern Languages, University of Helsinki.

Corpus of Late 18c Prose. Compiled by David Denison and Linda van Bergen, University of Manchester. hdl.handle.net/20.500.12024/2468.

Denison, David. 1994. "A Corpus of Late Modern English Prose." *Corpora Across the Centuries: Proceedings of the First International Colloquium on English Diachronic Corpora, St Catharine's College Cambridge, 25–27 March 1993*, edited by Merja Kytö et al., Rodopi, pp. 7–16. Language and Computers – Studies in Practical Linguistics 11.

Gardner, Anne-Christine. 2023a. "Petitioning for the Education of the Poor: The Sociohistorical Context of Self-corrections in a Late Modern English Draft Letter." *Intra-Writer Variation in Historical Sociolinguistics*, edited by Markus Schiegg and Judith Huber, Lang, pp. 157–180.

---. 2023b. "Speech Reflections in Late Modern English Pauper Petitions." *English Language & Linguistics*, vol. 27, no. 3, pp. 491–516.

---. 2023c. "English Pauper Letters in the Eighteenth Century and Beyond: On the Variability and Evolution of a New Text Type." *Sociocultural Change and the Development of Vernaculars in Early Modern Europe*, edited by Oliver Currie. Special Issue of *Linguistica*, vol. 63, no 1–2, pp. 301–336.

Gardner, Anne-Christine, et al. 2022. "Language and Mobility of Late Modern English Paupers." *Migrations and Contacts*, edited by Michael C. Frank and Daniel Schreier, Universitätsverlag Winter, pp. 45–70. Swiss Papers in English Language and Literature 41.

Hernández-Campoy, Juan Manuel. 2021. "Corpus-Based Lifespan Change in Late Middle English." *Language Variation and Language Change across the Lifespan: Theoretical and Empirical Perspectives from Panel Studies*, edited by Karen V. Beaman and Isabelle Buchstaller, Routledge, pp. 164–181.

Hernández-Campoy, Juan Manuel, and Tamara García-Vidal. 2018. "Style-shifting and Accommodative Competence in Late Middle English Written Correspondence: Putting Audience Design to the

Test of Time." *Folia Linguistica Historica*, vol. 39, no. 2, pp. 383–420.

King, Steven. 2009. "'I Fear You Will Think Me Too Presumtuous in My Demands but Necessity Has No Law': Clothing in English Pauper Letters, 1800-1834." *International Review of Social History*, vol. 54, no. 2, pp. 207–236. *JSTOR*, www.jstor.org/stable/44583131.

---. 2019. *Writing the Lives of the English Poor, 1750s-1830s*. McGill-Queen's UP.

Kytö, Merja. 2012. "New Perspectives, Theories and Methods: Corpus Linguistics." *English Historical Linguistics: An International Handbook*, edited by Alexander Bergs and Laurel J. Brinton, vol. 2, De Gruyter Mouton, pp. 1509–1530.

Kytö, Merja, et al. 2011. *Testifying to Language and Life in Early Modern England: Including a CD-ROM containing An Electronic Text Edition of Depositions 1560-1760 (ETED)*. John Benjamins.

Laitinen, Mikko, and Anita Auer. 2014. "Letters of Artisans and the Labouring Poor (England, c. 1750-1835): Approaching Linguistic Diversity in Late Modern English." *Contact, Variation and Change in the History of English*, edited by Simone E. Pfenninger et al., John Benjamins, pp. 187–212.

Lowth, Robert. 1762. *A Short Introduction to English Grammar*. A. Millar, R. and J. Dodsley.

Nelson, Mike. 2010. "Building a Written Corpus. What Are the Basics?" *The Routledge Handbook of Corpus Linguistics*, edited by Anne O'Keeffe and Michael McCarthy, Routledge, pp. 53–65.

Nevalainen, Terttu, and Helena Raumolin-Brunberg. 2017. *Historical Sociolinguistics: Language Change in Tudor and Stuart England.* 2nd ed., Longman.

Rissanen, Matti, et al. 1991. "The Helsinki Corpus of English Texts." *ICAME Collection of English Language Corpora* (CD-ROM), 2nd ed., edited by Knut Hofland et al., The HIT CENTER, University of Bergen, Norway.

Sokoll, Thomas, ed. 2001. *Essex Pauper Letters, 1731-1837*. Oxford UP.

Stratton, James. 2020. "Corpora and Diachronic Analysis of English." *The Routledge Handbook of Corpus Approaches to Discourse Analysis*, edited by Eric Friginal and Jack A. Hardy, Routledge, pp. 202–218.

The Bluestocking Corpus: Private Correspondence of Elizabeth Montagu, 1730s-1780s. 2017. First version. Edited by Anni Sairio, XML

encoding by Ville Marttila. Department of Modern Languages, University of Helsinki. bluestocking.ling.helsinki.fi

The Mary Hamilton Papers (1743-1826). 2019-2023. Compiled by David Denison, Nuria Yáñez-Bouza, Tino Oudesluijs, Cassandra Ulph, Christine Wallis, Hannah Barker and Sophie Coulombeau, University of Manchester. www.digitalcollections.manchester.ac.uk/collections/maryhamilton.

Whyte, Ian. 2004. "Migration and Settlement." *A Companion to Nineteenth-Century Britain*, edited by Charles Williams, Blackwell, pp. 273–286.