Year : 2022

Exploring gendered discourse in late 18th- and early 19th-century pauper letters from the LALP project: the importance of the corpus

Marina Berts

UNIL | Université de Lausanne
Faculté de théologie
et de sciences des religions

UNIL | Université de Lausanne
Faculté des lettres

UNIL | Université de Lausanne
Faculté des sciences sociales
et politiques

Mémoire de Maîtrise universitaire interfacultaire en Humanités numériques

Exploring gendered discourse in late 18th- and early 19th-century pauper letters from the LALP project: the importance of the corpus

Présenté dans la discipline « Anglais »

par

Marina Berts

sous la direction de la Professeure Anita Auer

et la codirection du Professeur Michael Piotrowski

Session de juin 2022

Exploring gendered discourse in late 18th- and early 19th-century pauper letters from the LALP project: the importance of the corpus

Abstract

Historical ego-documents and pauper relief requests are important research objects in historical sociolinguistics as they represent non-standard language of the past and language history 'from below'. Today, many available electronic historical corpora exist, and interdisciplinary research allows for new research angles and research questions to develop, for example gendered discourse. To investigate a potential existence of gendered discourse in the LALP corpus of pauper relief requests, a subcorpus of 20 pauper letters – ten letters written by female and ten by male paupers – was created. The results of the research analyses indicate that the lack of schooling and access to standard English influenced on the written language of relief applications substantially: letters written by both genders present phonetic spelling, absence of punctuation, erratic capitalisation, and a preference for Anglo-Saxon monosyllabic words. Nevertheless, paupers seemed to be familiar with the Latinate vocabulary connected with letter writing and the rhetoric of relief requests. Female paupers seem more inclined than men to use a more revered and subdued linguistic attitude when addressing themselves to the officials. Male paupers often have recourse to direct threats when applying for relief while women 'bake in' their threats in the narration. Male paupers also use more varied *that*-clauses than female paupers. In spite of minor linguistic differences found in the relief applications written by female and male paupers, no convincing evidence of gendered discourse could be found.

Contents

Chapter 1

Introduction

In the 20th century, gendered discourse became a field of interest to linguists and sociolinguists, and the influence of social factors on language was much discussed, especially in the second half of the century (Weinreich, Labov & Herzog 1968; Labov 1972; Lakoff 1973; Romaine 1982; Chambers 1995; Biber & Burges 2000; Labov 2001 and many more). Differences in modern-day spoken and written language between women and men continue to be researched from different angles, often in connection with automated tasks performed by machines (Blaxter 2014; Bolukbasi et al. 2016; Menegatti & Rubini 2017). Today, in the 21st century, there is on the whole more information and theories on gendered discourse than earlier. But not only contemporary speech and texts are investigated by linguists and sociolinguists today. Documents from earlier time periods are also being used as research material on gendered discourse (Nevalainen 2000; Nevalainen 2002; Daybell 2006; Nevalainen & Raumolin-Brunberg 2017). Even though there are no records of spoken English for time periods such as the Early and Late Modern English periods (at least not for the period up to the the end of the 19th century), speech-like texts exist from those days, for example court proceedings and ego-documents such as personal letters (Daybell 2006; Elspaß 2012; Auer et al. 2014). These historical documents, evidence of what is called 'language from below', provide us with traces of what the spoken language of the members of the lower social classes could have sounded like in the past. These texts are also rare fragments of the living conditions of artisans and paupers expressed in their own voices.

Investigating gendered discourse in pauper relief requests written during the last four decades of the English Old Poor Law (ca. 1795-1834) raises many other important questions in connection with the language of the lower social classes in the Late Modern English period, such as schooling and literacy. As we do not have access to spoken language from that period, it limits historical sociolinguistic investigations to written records. Pauper relief requests present speech-like features, as the letter writers were only partly schooled and thus used much phonetic writing. And yet, these letters are not authentic speech situations like those that we have access to today (Smitterberg 2012; Auer et al. 2014). Nevertheless, these letters represent 'written material that reflect spoken language as closely as possible' (Auer et al. 2014: 9). As Smitterberg (2012: 954) argues, '[t]he fact that large segments of the population were thus unable to produce written texts makes it more difficult to obtain language data representing some sociocultural groups than is the case for Present-day English'. Therefore, pauper relief requests are very interesting documents for language studies, as they represent the words and writings of party-schooled people who 'were expected to copy others' words, not write their own' (Fairman 2011: 41) and were not used to writing.

The study of social factors surrounding the language of pauper letters revealed that schooling does not represent only a small detail in the given time period. Up to the end of the 19th century, education was the privilege of members of the upper social classes that could afford to pay for classical education. The children of the lower social classes attended such schools that were free, for example charity schools, dame schools or Sunday schools (Purvis 1989; Fairman 2011), if they had a chance to attend at all. We know that women and men received different kind of schooling in the middling and elite layers of society (Lawson & Silver [1973] 2013). However, it is not known if this holds for the members of the lower classes. Literacy had increased in the 18th century (Smitterberg 2012: 953), but in the following century, working-class women were still only reluctantly admitted to for example Mechanic's Institutes (Purvis 1989). As we shall see later on in this *mémoire*, women were also subject to patriarchal ideologies that kept them in a specific social role (see 2.3).

The factors to take into account in order to analyse pauper relief letters being numerous, I asked myself how best to approach my research question. Would a qualitative analysis give me the answers to the search for possible gendered discourse in pauper

letters? As corpus linguistics is much used in both linguistics, sociolinguistics and historical sociolinguistics, I opted for this method to obtain additional results not available through a qualitative analysis. As a student of Digital Humanities and English linguistics, my next query was how to approach the question of the corpus. I therefore decided to dedicate some space in this paper to the importance of the (electronic) corpus in linguistic research.

The pauper relief requests raise a number of questions related to linguistics and sociolinguistics, e.g the concept of literacy, traces of standard English that was being codified in the 18th century as well as possible differences in vocabulary and/or grammar between female and male paupers. As '[i]n stable situations, women perceive and react to prestige or stigma more strongly than men do, and when change begins, women are quicker and more forceful in employing the new social symbolism' (Labov 2001: 291), could this also be the case in the 18th and 19th centuries? And what about Labov's Gender Paradox, that indicates that '[w]omen conform more closely than men to sociolinguistic norms that are overtly prescribed, but conform less than men when they are not' (Labov 2001: 293)? This concerns contemporary language in a context where standard written English is more the rule than the exception, and agreeing with Nevalainen (2018: 6), I am not certain of 'the extent to which the methods and models used in modern sociolinguistics could be applied to the past'. Could the Gender Paradox also have been relevant in the late 18th and early 19th centuries when the English language was being standardised and codified? If we take into consideration the Uniformitarian Principle (UP) that states that ongoing processes today may well be due to similar processes in the past (Bergs 2014), even though this principle is not unproblematic (Trudgill 2020), it would be possible to presume that women and members of the lower classes also drove language change 200 years ago. Therefore, I presume this could be detected in the language of female and male paupers in their relief requests.

Within the context outlined, the aim of the current *mémoire* is to approach any apparent differences between female and male paupers on a more general level. The linguistic elements to investigate are phonetic writing, punctuation, traces of standard writing, vocabulary and the opening and final salutations of the pauper letters. The creation of a subcorpus of 20 pauper letters, ten written by female paupers and ten by male paupers, also proved necessary in order to proceed to a quantitative analysis. This of course

led me to the spelling standardisation of the letters present in my subcorpus, process that proved to be quite challenging. I suspect my *mémoire* project will raise more questions than give answers, but I hope it will give clues and indications to further investigations of possible gendered discourse in pauper letters.

I have structured my *mémoire* going from the more general questions around the corpus to the more precise, from the broader notion of the corpus to the analysis of a subcorpus of pauper letters. In the first section of Chapter 2, I will discuss the role of the corpus as a tool and as a model. I will start with the definition of the corpus and how it relates to the formal model according to Digital Humanities to continue with the processes scholars are facing when wanting to make a hand-written historical corpus machine-readable. These steps are crucial in order to produce texts in standard English that will allow for a computational quantitative analysis of the material. The questions of uncertainty and bias related to corpus design will then be discussed as well as the 'bad data' problem intimately linked with historical documents that enter a corpus. In the second section of Chapter 2, I will describe what pauper letters are, discuss their main characteristics as well as their importance for the study of language use and language change. I will also present the LALP corpus that is currently under construction. The last section will be devoted to a historical overview of women's lives in the late 18th and early 19th centuries, to present-day theories on gendered discourse and to linguistic differences between women and men.

Chapter 3 will give explanations as to the methodology I used to analyse the pauper letters. As a complete analysis of the LALP corpus was not conceivable for this *mémoire*, I decided to create a subcorpus of ten female and ten male pauper letters from the LAP corpus. The subcorpus was then used to make a comparison of the language in these letters. Chapter 4 presents first the results of a qualitative analysis, then the spelling normalisation process and finally the quantitative analysis. My qualitative analysis examines the address form and the final salutations, lexical differences, punctuation, traces of standard writing and oral features. In the subsection on spelling normalisation, I will explain the challenges that I was faced with when using the program VARD2. The quantitative analysis is composed of word counting, word lists, the investigation of *that*-clauses and the analysis of vocabulary. Finally, I will explore the potentiality of making a hypothesis on gendered discourse in the LALP pauper letters. Chapter 5 will be dedicated to a

discussion on my findings and I will end this *mémoire* with a conclusion in Chapter 6. The Appendix gives the codes of the pauper letters that constitute my subcorpus and the English archives where these letters can be found.

Chapter 2

Exploring gendered discourse in late 18th- and early 19th-century pauper letters from the LALP project: the importance of the corpus

This section is divided into three parts and serves primarily as a general background to the exploration of gendered discourse in late 18th- and early 19th-century pauper letters from the LALP project and to the importance of the corpus.

In the first subsection (2.1), I will consider more general questions of the corpus as a tool in historical sociolinguistics. I will discuss the definition and characteristics of a corpus in general and what a corpus represents in Digital Humanities from a more theoretic point of view. The corpus as a model in historical sociolinguistics will then be treated, as well as the question of making a historical corpus machine readable, allowing thus a more profound exploration of the corpus material, rendering possible a combination of qualitative and quantitative research. The two last parts of this subsection will deal with uncertainty and bias on the one hand and the 'bad data' problem on the other.

The second part of this section (2.2) is dedicated to pauper letters and the LALP corpus. First, I will present the relief letters and I will also explain the main characteristics of these relief letters before continuing with the importance that these letters

represent for the study of Late Modern English language use of the lower classes and language change. I will also describe the SNSF-funded LALP corpus project.

The last part of this section (2.3) will deal with the question of finding potential evidence of gendered discourse in the LALP relief letters. By looking into sociolinguistics, history and contemporary research on language and gender, I will discuss the importance of approaching the relief letters from the angle of historical sociolinguistics.

2.1 The corpus as a tool

A corpus is a powerful tool in research, especially today as it is possible to work with electronic corpora. The digitisation of historical documents and the application of new digitalisation processes have allowed for new research questions to emerge. Machine-readable corpora have been widely used in corpus linguistics (CL) for several decades now, and the rise of Artificial Intelligence (AI) may yet contribute even more in making CL research progress. But what is a corpus and what are its pitfalls? I will now examine these questions by inspecting the corpus both theoretically and practically in the next subsections. I will start by discussing the definition and characteristics of the corpus (2.1.1) and link the corpus to the field of Digital Humanities (2.1.2). Then, I will focus on the creation of a digitised corpus as a model in historical sociolinguistics (2.1.3 and 2.1.4) and finally present possible pitfalls and problems tied to corpora (2.1.5 and 2.1.6).

2.1.1 The corpus – definition and characteristics of a corpus

The word 'corpus' is Latin meaning 'body' (*OED online* 2021), and the definition of a corpus is 'a body of naturally occurring language' (McEnery, Xiao & Tono 2006: 4, as quoted in Gries & Newman 2013: 258). Another perhaps more exhaustive definition of the linguistic corpus is 'a collection of written or spoken language material in machine-readable form, assembled according to precise criteria and for the purpose of studying one or more specific linguistic phenomena' (Piotrowski 2019b: 12). Linguistic corpora are generally contained in a data base and are thus machine readable, which allows researchers to use corpus linguistics methods to explore different aspects of language from various angles in a minimal lapse of time (Cantos 2014). As corpora are to be used for

different research questions, they take disparate forms, representing various media types, such as written texts (in both digital and non-digital form), audio recordings and their transcriptions, videos combined with audio and so on (Gries & Newman 2018: 258). In other words,

> the prototypical corpus is a *machine-readable* collection of language used in authentic settings/contexts: one that is intended to be *representative* for a particular language, variety, or register [...] and that is intended to be *balanced* such that the sizes of the parts of the corpus correspond to the proportion these parts make up in the language/variety/register. (Gries & Newman 2018: 258)

A corpus can thus be of different types, depending on the research question: if we are exploring a specific variety of language in a certain context, it will be necessary to create a corpus containing only texts representative of that language type. On the other hand, if we are aiming for a broader approach to language, it will be indispensable to include as many text types from as many periods as possible, for example a large collection of texts from the Internet. This is what Partington (2006: 2nd par.) calls heterogeneric and monogeneric corpora: a corpus may be vast, containing millions of words and representing many different genres (heterogeneric), or it can be much smaller if we are seeking to create a specialised corpus for which there are only a few reference texts representing one specific text type (monogeneric).

There are other criteria to respect when creating a corpus of texts apart from the balancedness and representativeness mentioned by Gries & Newman (2018): there is also a need for document authenticity. It is crucial that '[the] texts involve authentic cases of language use as it occurs in the "real world", as opposed to say a made-up sentence by a linguist in order to demonstrate a particular point' (Baker 2014: 7). A corpus should thus reflect 'real' language as it is used today, or if the corpus contains historical documents, as language was used in the past. The combination of these three criteria, balancedness, authenticity and representativeness, thus become essential in corpus creation: 'a corpus needs to be sampled in such a way that we can be confident that it actually is representative of the language variety that we are studying' (Baker 2014: 7). If a corpus is to be the tool for exploring the language of Afro-American students between 12 and 16 years of age, it should contain examples of speech and texts from such a group and not from any other

body, such as for example middle-aged CEOs in multinational companies (cf. Nevalainen 1999: 502). One of the most important challenges of building a corpus is respecting these criteria, because it is essential 'to minimize distortion in the mapping from the original (the language) to the model (the corpus)' (Piotrowski 2019b: 12). Creating distortion and perhaps even bias when building a corpus will inevitably influence on the results of the subsequent corpus analysis.

What then is the difference between a collection of texts and a linguistic corpus? It could be argued that there is no clear demarcation between the two. As already mentioned, a corpus (as well as a collection) usually contains large quantities of data, and these data can be analysed in different ways depending on the research question, especially today since it is possible to explore enormous amounts of native digital or digitised texts computationally. The Web is often used as a corpus even though it lacks representativeness or metadata. A small collection of historical texts might serve as a corpus for research in spite of its comparatively reduced size. Gries & Newman (2018: 258) claim that '[b]eyond being a body of naturally occurring language, then, it is difficult to agree on any more particular definition of what a corpus is or is not'. So could the distinction between a collection and a corpus be their size, or could the difference lie elsewhere? According to Piotrowski (2019b: 12), a corpus distinguishes itself from a collection of texts because the texts of the corpus have been carefully selected according to the criteria of the research question and they are in a form that is machine readable. This suggests that a corpus does not correspond to just any loose collection of texts. The corpus functions as a model of a specific object and has been carefully built for a particular purpose and for a particular public. Linguistic corpora are therefore to be considered as formal models of language 'since they exhibit all three of Stachowiak's model properties, the mapping property, the reduction property, and the pragmatic property' (Piotrowski 2019b: 14)[1]. Contrary to the collection, the corpus, being a model of an original, maps the attributes of the original texts while reducing the number of these attributes: the model replaces and represents the original.

Corpora might not necessarily always be machine readable even though most nowadays are. Quantitative analyses are much more difficult to perform on documents and texts that are not machine readable, which means that much effort has been put

[1] see 2.1.2 on Digital Humanities and modelling

into the electronic conversion of hand-written documents. Transforming historical documents, especially hand-written ones, into a machine-readable format nevertheless requires considerable time and efforts as they call for extensive manual work, such as transcription and encoding. As technology has advanced, the number of historical electronic corpora has considerably increased over the years (López-Couso 2016: 127–129). Today, there is access to both corpora that are digitally native and (in the case of historical linguistics) corpora that have been digitised[2]. Apart from the LALP corpus which is currently being created and is described below (2.2.4), there are other historical corpora in existence such as the *ARCHER Corpus* (ongoing project that started in 1990), the *Helsinki Corpus of English Texts* (1991), the *Corpus of Early English Correspondence Sampler* (CEECS, 1998), the *Eighteenth Century Collections Online* (ECCO, 2004), the *Corpus of English Dialogues 1560-1760* (CED, 2006), the *Parsed Corpus of Early English Correspondence* (PCEEC, 2006), and the *Helsinki Corpus of Scottish Correspondence (1540-1750)* (2021), just to mention a few corpora available in English.

But the digitised corpus is not, however, the panacea to all research questions. The digital format opens up new possibilities for corpus analysis through corpus linguistics, but it can also lead researchers to what Piotrowski calls the 'allure of convenience' (Piotrowski 2019b: 15). Easy access to machine-readable corpora might encourage scholars and scientists to 'opportunistically choosing a particular periodical as a "model" for some phenomenon just because it is available as a corpus' (Piotrowski 2019b: 15). Not only periodicals might be chosen as a 'model' but any digitised corpus that is available and perhaps existing in open access might become a target for easy analysis. It is a fact that computer-assisted research and investigation can save much time and efforts, especially for repetitive tasks that can be a chore to any scholar. Nevertheless, only because access to computationally searchable corpora is much easier today than 20 or 30 years ago and the results of queries are produced rapidly, it does not mean that the careful planning and building of a corpus nor a rigorous interpretation of analysis results may be bypassed.

In the light of what precedes, the creation and use of a corpus is then not as straightforward as it might seem. There are important challenges that are to be taken into consideration when operating corpus design and analysing data, not only the allure of convenience. Biber (1990) discusses the main issues concerning the constitution of a

[2]see https://clarin.eu/resource-families/historical-corpora

corpus. First, Biber deals with the question of the appropriate length of the texts contained in a corpus – are long texts necessary in the corpus or are shorter texts just as viable for obtaining reliable results? Then, there is the issue of text categories: how do we define categories such as literary genre, for example? As Biber points out, the term 'genre' is not easily defined 'since texts within a genre can differ markedly in their linguistic characteristics' (Biber 1990: 261). In fact, the novel and the newspaper article do not correspond only to one type of linguistic expression but may contain different sorts of language. As an example, Biber highlights the case of the newspaper article that can take a colloquial linguistic form or follow a more elaborate and informational pattern. Finally, Biber mentions the overall size and composition of the corpus and concludes by stating that most existing corpora are perfectly acceptable and that it is therefore not necessary for researchers to limit themselves to large corpora:

> my purpose has been to show that existing corpora are adequate in many respects: in particular, that relatively short text lengths and small corpus size are often adequate, that genres are well-defined text categories, and that the design goal of representing a wide range of variation (adopted by existing corpora) is necessary if a corpus is to be used for analyses of textual variation. (Biber 1990: 269)

As long as the corpus is explicit, unambiguous and coherent, it will be possible to use it for linguistic analyses. In agreement with Biber (1990: 269), Gries & Newman (2018: 259) and Piotrowski (2019b: 12), it is thus possible to affirm that a corpus can be validated as reliable even though its size is small or contains shorter texts, provided that the corpus is well-constructed and well-balanced.

Historical corpora (and corpora in general) can contain not only texts but can also be amplified with supplementary information such as metadata, textual markup and annotations (McEnery & Hardie 2012; Piotrowski 2012). Metadata provides information about the text, for example the identity and gender of the writer, her or his age, what language was used and when the text was written. Textual markup is more about the text itself: it gives information about the formatting of the text, for example if certain parts of the text are titles, descriptions or text in italics (McEnery & Hardie 2012: 29). In speech transcriptions, the markup might also indicate where a speaker begins to utter a sentence and when the utterance finishes (McEnery & Hardie 2012: 29). An annotation can be a

tag or a label expressing for example word class or pronunciation. The annotation process uses a specific annotation system such as XML that 'encode[s] linguistic information within a corpus text in such a way that we can systematically and accurately recover that analysis later' (McEnery & Hardie 2012: 30). All this extra data are sometimes essential in order to obtain reliable results in line with the research question. If this information is absent, it might be challenging to produce results of corpus analysis, as there will be loss of information or even a problem with 'bad data' that could distort the final results (see 2.1.6).

As mentioned, above, a well-structured and reliable corpus thus corresponds to a model of the original texts contained in the corpus. But what then is a model? In the next subsection, I will continue the reflection on the corpus as a model and on the challenges encountered when modelling historical texts.

2.1.2 Digital Humanities (DH), models and modelling

In this subsection, I will first briefly present the field of Digital Humanities (DH) and then continue with the definition of the term 'model' and explain what modelling is. Finally, I will discuss the importance of model-making in DH.

Digital Humanities

As the importance of digitisation, i.e. transforming analogue information into a digital format, and digitalisation, the 'socio-technical processes surrounding the use of (a large variety of) digital technologies that have an impact on social and institutional contexts that require and increasingly rely on digital technologies' (Rijswijk et al. 2020: 1) increases rapidly everywhere in society today, it becomes important to implement these digital technological advances in the Humanities research as well. This trend has already taken root in certain fields such as linguistics (Auer et al. 2015: 6), and other fields of the Humanities are currently also going in the same direction. Even though somewhat unknown amongst the general public and in academia,

> the digital humanities today is about a scholarship (and a pedagogy) that is publicly visible in ways to which we are generally unaccustomed, a scholarship and pedagogy that are bound up with infrastructure in ways that are deeper and more explicit

> than we are generally accustomed to, a scholarship and pedagogy that are collaborative and depend on networks of people and that live an active 24/7 life online. (Kirschenbaum 2010: 60)

Digital Humanities (DH) is thus not only a question of using computational tools but also of interdisciplinary collaboration across boundaries. Just like historical sociolinguistics, DH has been developing for more than twenty years, and yet, there is still no consensus on a unique definition of DH. The general idea that currently surrounds DH is that this field simply introduces computers and computational tools into qualitative research in order to maximise the research process and also to obtain statistical data. Kirschenbaum (2010: 56) argues that 'digital humanities is more akin to a common methodological outlook than an investment in any one specific set of texts or even technologies'. Piotrowski claims that DH is a concept that generally 'refers to digitization of research objects in order to automate mechanical tasks and to apply quantitative methods at a larger scale' (Piotrowski 2019b: 9) but also that DH is much more than that (cf. Van Zundert 2016). DH is about 'the development and use of *new methods*' (Piotrowski 2019b: 9) in the Humanities. First of all, says Piotrowski, to meet the challenges of the digitised contemporary world, the Humanities need to adopt more agility and create computational models in order to be able to 'test them rigorously, publish them early for comments, and iteratively and incrementally improve them by integrating feedback from testing and from other scholars' (Piotrowski 2019b: 10). According to Piotrowski, computational models are therefore essential to DH as they represent a central role in the development of the Humanities.

Taking the above into account, what then could be a relevant definition of Digital Humanities? Piotrowski argues that as society is transformed by digital technologies, Humanities 'can meet their scholarly challenges and societal responsibilities only when the digitalization of the humanities is understood as an actual transformation' (Piotrowski 2019b: 10). The key to such a transformation by digitalisation is not only essential for the Humanities but also for society as a whole. And yet, even though the understanding of generalised digitalisation seems clear, the definition of DH is much disputed. In Piotrowski's definition of DH, two important aspects representing the idea of duality are to be found: DH are (i) 'research on and development of means and methods for constructing formal models in the humanities (*theoretical Digital Humanities*)' and (ii)

'the application of these means and methods for the construction of *concrete* formal models in the humanities disciplines (*applied Digital Humanities)*' (Piotrowski 2019b: 10). Applied DH is current in academia today, but theoretical DH still has room for further development. It is important to take into account that models are to be adapted to the different fields of the Humanities, which means that these computational models will vary from one research question and research field to another and would therefore benefit from theoretical DH. As Pavé (2005: 170) underlines, the model is not stable and « il ne faut pas hésiter à faire évoluer le modèle constamment quand la situation l'exige » ('one should not hesitate to make the model evolve whenever the situation requires it', my transl.). The aim is therefore not only to make models evolve and develop, but also to create new methods of research for specific purposes (see Domingo & Casacuberta 2020).

Models and modelling

Just as the definition of DH is still widely debated, the terms 'model' and 'modelling' are also being discussed. Every day, we – as humans – create mental models based on our environment and of all that surrounds us (Piotrowski 2019b: 10). Without mental models, it would be difficult to create a coherent image of what is generally believed to be reality. So if model making is such a basic human activity, what is the definition of a model? There is a multitude of definitions of what a model is, and I will illustrate some of them here.

When discussing the model, Sauret (2017: 5–6) emphasizes that there are two main interpretations of the term 'model': the first interpretation corresponds to a simplified representation of an object that does not contain all the properties or details of the original object, for example a small-scale model of a ship or a car. The second interpretation is the opposite: a theoretical model is first created and the object is subsequently constructed in order to test the model. Sauret thus asserts that the model carries a dual substance as it places itself between the real object and its representation. Sauret argues that this duality is known in data science, where we find « les modèles descriptifs (du réel vers le modèle) et les modèles prédictifs (du modèle vers le réel) » ('descriptive models (from reality to the model) and predictive models (from the model to reality)', my transl.) (Sauret 2017: 6). This seems to indicate that the model has a nature of its own,

being a representation of an object but still existing somewhere between reality and embodiment. According to McCarty (2002: 104), 'a model is a manipulable knowledge representation', definition that seems to generate the same idea as Sauret's of a representation that does not entirely correspond to the original but that can be altered and developed. Meunier (2017: 26) gives a precise definition of the term 'model' as he claims that « [d]ans une pratique scientifique, un modèle est un artefact sémiotique servant à décrire, expliquer et comprendre des objets » ('in scientific practice, a model is a semiotic artefact whose aim is to describe, explain and understand objects', my transl.).

These definitions seem to suggest that the aim of a model is to describe a given phenomenon or object, to explain its significance and behaviour and to give an explanation, or at least to formulate a deduction, of the function and role of the original phenomenon or object. Basing his description of the model on Stachowiak's general model theory, Piotrowski (2019b) also supports the idea of duality as he describes the (formal) model and modelling in the following terms:

> The basic assumption is that arbitrary objects can be described as *individuals* characterized by a finite number of *attributes*. Attributes can be characteristics and properties of individuals, relations between individuals, properties of properties, properties of relations, etc. Modelling is then a mapping of attributes from the original to the model. (Piotrowski 2019b: 11)

Piotrowski thus argues that the model can therefore be defined through three main properties, according to Stachowiak's theory (Piotrowski 2019b: 11):

1. a model is based on originals that can themselves be models of something (mapping property) – the model thus 'maps' different individuals.

2. the model of the individuals will reiterate at least some of the attributes of the originals but not all (reduction property). Those who create and use the model decide what attributes are represented in the model.

3. a model has a replacement function as it is created for someone (a machine or a human), carries a specific purpose (a reason for being created), and is generally used for a specific amount of time (pragmatic property).

A model is accordingly a map of objects that combines certain chosen attributes of the original, reflecting a sufficiently clear 'map image' of the original individuals and their attributes that can consist of characteristics, properties and relations. Ciula & Eide (2017: 38) discuss this same idea and argue that 'by modelling we link models to qualities and relationships already existing in the objects being modelled. Such linking is based on choices which are made for a certain end informing and motivating the act of modelling'. Modelling thus depends heavily on the choices and intentions of those who operate the modelling process. In the case of corpora, all three properties of Stachowiak's model theory are present: a corpus maps individuals, it reflects certain attributes according to the choices of the model creators, and its purpose is to be used in research by different specialists over a certain lapse of time.

Considering the previous points made, it could then be argued that the model is not an exact replica of the original. Piper (2017: 652) focuses on the main concern of the model, that of its representational qualities, as he claims that 'models shift the focus toward the signifiers of research and away from the signifieds'. Here, Piper draws the attention to the fact that models are not the original objects themselves and should not be mistaken for such (see also Piotrowski 2019b). And yet, it is necessary for a model to be well-formed, well-balanced and based on rigorous information. This is what mental models or verbal theories are not: they need therefore to be written down and 'mapped' in order for them to become reliable models, because 'language is inherently (and adaptively) vague and ambiguous' (Smaldino 2020: 207). As both Piper and Piotrowski also argue, the whole object does not correspond to its model, which entails that models need to 'simplify and hence ignore many of the nuanced details of the real world' (Smaldino 2020: 207). This kind of 'omission' does not need to be problematic but could be regarded more like 'a feature rather than a bug' (Smaldino 2020: 207). Because of such simplification, it is essential to test the model and correct it if necessary, as well as being aware of its limitations and take these into consideration when drawing conclusions and interpreting results[3].

Often, the term 'formal model' is used in model-making, in contrast to mental models. According to Piotrowski (2019b: 12), the word 'formal' refers to the fact that the model needs to be unambiguous, explicit and leave no room for incoherence. Com-

[3]for a brief presentation of models on Youtube, see https://bit.ly/3nUmoMl

puters are therefore the perfect tools for dealing with formal models. The elaboration of models that show no signs of ambiguity and that are explicit and coherent is essential in all research as it will allow any researcher to implement the formal model and repeat the experiments based on the models so as to confirm the exactitude of the model. This is also one of the sixteen reasons that Epstein (2008: par. 1.5) mentions for modelling: 'In *explicit* models, assumptions are laid out in detail, so we can study exactly what they entail. [...] By writing explicit models, you let others replicate your results'. Questioning results and repeating the analysis process allows scholars to either confirm or invalidate the model and the theories or hypotheses based on it. Model-making is therefore extremely useful, and it actually already exists in the Humanities and has done so for several decades. As Biber, Conrad & Reppen (1996) have already pointed out, model-making was largely used in the linguistic field as early as the 1990s, especially in corpus linguistics. Access to electronic corpora allows most linguistic research to combine both qualitative as well as quantitative aspects of the research material and to present extended findings based on this joint process. Not unexpectedly, other fields such as historical sociolinguistics also benefits from document digitisation and combined analyses (Kytö 2011; Tumbe 2019).

The elaboration of formal models in the Humanities does not only evoke the combination of qualitative and quantitative methods but also raises another widely debated question. This question is linked to computational methods and modelling and could be formulated like so: Does creating and writing formal models necessitate the skills to code them? Coding seems to be a part of DH and model-making (Van Zundert 2016) as Humanities and Social Sciences scholars are encouraged to '[g]rab some model code or jot down some relational equations and start messing around' (Smaldino 2020: 208) with NetLogo or ready-made code on GitHub. It is probable that researchers from the Humanities already have some coding knowledge, and in interdisciplinary projects, this knowledge will automatically augment spontaneously through contact with specialists in the Computational linguistics or Data science fields. Specific Digital Humanities programs in universities and other institutions will also enhance the computer skills of the students and give better access to computational thinking. The necessity for Humanities scholars to know how to code is, however, not a question that will be treated in depth in this *mémoire* but the subject would merit further discussion.

Considering the various definitions of the term 'model', I will adhere to Piotrowski's definition (Piotrowski 2019b: 11) in this *mémoire* and will, in the next subsection, discuss the corpus as a model in historical sociolinguistics.

2.1.3 The corpus as a model in historical sociolinguistics

This subsection will present different aspects of the corpus as a model in historical sociolinguistics. First, I will start with a brief introduction to what historical sociolinguistics is and then continue with a discussion around the importance of corpus design of historical texts, before treating other questions such as how to make a corpus machine readable, how to approach uncertainty and bias, and finally how to deal with the 'bad data' problem.

Historical sociolinguistics has developed over the last three decades as a subfield of linguistics. As early as 1968, Weinreich, Labov & Herzog already introduced the idea of sociohistorical linguistics:

> With their emphasis on the need to incorporate external factors into a theory of language change and to transcend the old dichotomy of synchrony and diachrony, Weinreich et al. (1968) laid the foundations for an approach to language that was inherently historical and social. As such, this paper marked the emergence of a new field of inquiry, viz. sociolinguistics, which has been expanding ever since. (Auer et al. 2015: 2)

In 1982, the term 'socio-historical linguistics' was used by Romaine (1982), and an alternative term, 'historical sociolinguistics', has later on been employed by other scholars, amongst others Milroy (1992). Ever since, there has been an important development of the field which has led to the establishment of its basic assumptions not only for the English language, but also for languages such as German and Dutch. These efforts were crowned by the foundation of the Historical Sociolinguistics Network[4] that gather those scholars and lay people who have an interest in historical sociolinguistics (for a more detailed overview of the field, see Auer et al. 2015). Other recent milestones in historical sociolinguistics are the *Journal of Historical Sociolinguistics*[5], the Blackwell *Handbook*

[4] http://hison.sbg.ac.at/

[5] https://degruyter.com/journal/key/jhsl/html

of Historical Sociolinguistics (Hernández-Campoy & Conde-Silvestre 2014) and Robert McColl Millar's course book *English Historical Sociolinguistics* (2012), proving that the interest in the field of historical sociolinguistics is growing.

The aim of historical sociolinguistics is 'to study language use, as produced by individual language users, embedded in the social context in which these language users operate, and understood not only from a communicative angle but also as conscious or unconscious acts of identity and social distinction' (Auer et al. 2015: 9). Researching historical documents thus requires adopting an extensive view of the time period, exploring not merely linguistic issues but also other important factors in connection with language such as social status, gender, identity, class belonging, prevailing ideologies, economical issues and societal phenomena such as migration. Drawing on different fields, historical sociolinguistics is thus a multidisciplinary field:

> historical sociolinguistics is by its very nature a multidisciplinary endeavour, drawing heavily on advances in social and cultural history, philology and paleography, corpus linguistics and modern-day sociolinguistics, as well as sociology and social psychology – even more so than in more traditional approaches to historical linguistics. (Auer et al. 2015: 8)

In addition to the multidisciplinary aspect of historical sociolinguistics, the advances in digitisation and computational sciences add a new layer of research opportunities for historical sociolinguistics. Different sources that can prove to be very useful or even essential to historical sociolinguistics research, such as historical dictionaries, writing manuals, guides of all sorts and other collections of texts from the corresponding time period, are today more accessible since they are to be found in digital form. This entails that several textual resources can be more easily combined than was the case earlier:

> Although English historical linguistics has always been heavily anchored in textual evidence, the last three decades or so have witnessed an increasing interest in the compilation of structured and systematic collections of texts from earlier periods of the language, mostly in computerized form. The availability of 'old material' in new formats, including not only electronic corpora, but also electronic dictionaries and online collections of texts, which provide quick and easy access to a large amount and a wide variety of data, has undoubtedly stimulated new research methods and

> approaches [...] and has enabled scholars to ask new questions and to reconsider old questions in a different light. (López-Couso 2016: 127)

Multidisciplinarity is thus essential for research in historical sociolinguistics, and having access to a variety of different sources is a necessity. Digitised historical documents give the researchers the opportunity to combine sources as well as using the data for both qualitative and quantitative analyses. This enhances the central concerns of historical sociolinguistics that are primarily 'how and when changes are transmitted from one speaker to another, how new forms become established in speech communities, across age groups, professions or social strata, and how prestige, norms of correctness and speakers' attitudes toward specific forms may affect changes' (Auer et al. 2015: 4). Today, diachronic studies are made more widely available through access to digitised historical documents worldwide. Earlier, consulting such documents was practically impossible or at least extremely time-consuming and even expensive. Language change can thus be studied more extensively by the use of digitised material.

It seems however clear that research in historical sociolinguistics faces challenges that do not necessarily exist for present-day language. The further the researched time period is from us time-wise, the more difficult it becomes to find the relevant data, such as metadata and other facts that surround spoken language and language use. This is what is commonly called the 'bad data' problem which will be discussed in subsection 2.1.6. But is it not better to use the data available than not using it at all? Historical sociolinguistic research has its limitations if we compare it to the bulk of data readily available today, and yet,

> [j]ust because there are methodological problems, we should not consider [historical sociolinguistics] an empirically invalid and inaccurate field of research. It is crucial in those areas of study for which oral records are not available, especially when studying long-term developments in language variation and change. (Hernández-Campoy & Schilling 2014: 74)

Exploring linguistic data from a historical sociolinguistics point of view can prove to be profitable if we rely on the Uniformitarian Principle that states that 'knowledge of processes that operated in the past can be inferred by observing ongoing processes in the present' (Christy 1983: ix, as quoted in Labov 1994: 21). The Uniformi-

tarian Principle thus presumes that the kind of language change and language variation that occurs today probably occurred in the same way in the past (Bergs 2014). Applying the Uniformitarian Principle strictly might, however, create anachronisms if scholars use patterns and processes adapted to contemporary language (Auer et al. 2015). Historical sociolinguistics is not only a question of language use and language change but it also contributes to 'reconstruct a broad picture of the social context in which the language varieties under investigation were used' (Auer et al. 2015: 5), which means that if we use our 21st century glasses to observe the past, we might produce anachronisms and bias. Nevalainen & Raumolin-Brunberg (2017: 6) also agree on this: '[i]t is obvious that present-day intuitions will not serve as secure guidelines for interpreting historical data in social terms'. When exploring language use and change that occurred in the past, it is necessary to adjust the contemporary lens through which we contemplate times gone (Bergs 2014). It is therefore important to use corpus data and corpus linguistics with caution, taking into consideration multiple factors linked to the period under research, for example socio-economical and socio-historical factors in connection for example with gender. One of the important concerns of historical sociolinguistics is 'to overcome the social bias connected to class, education and literacy inherent in written sources that has afflicted language historiography' (Auer et al. 2015: 6). This can be done efficiently if social factors are taken into consideration when interpreting for example quantitative analysis results.

Corpus design (and consequently model-making) based on historical documents should therefore be approached with care. As discussed in subsection 2.1.2, the model should leave no room for ambiguity nor incoherence and it should also be explicit. If the corpus is to be a model for studying historical language variation and change, the corpus needs to be designed in such a way that it corresponds to the criteria of the research question. Historically, writing was mostly done by an elite for official purposes and texts were often intended for science, politics and literature (Auer et al. 2014). Therefore, 'there will be more writing available from the middle and upper groupings of society since they are more likely to be (fully) literate, better able to afford quite costly items like paper and ink and capable of storing written material they have received' (McColl Millar 2012: 38). But even historical documents emanating from the elite have not been systematically conserved as '[they] survive by chance, not by design, and the selection

that is available is the product of an unpredictable series of historical accidents' (Labov 1994: 11). The conservation of written documents from days gone has thus often been unpredictable and inconsistent. As a consequence, it could be argued that building a corpus and using it as a model of a historical linguistic phenomenon seems to rely partly on arbitrary circumstances. In corpus design, and later on for corpus analysis, this remaining material and the results obtained should therefore be viewed critically and the aspects of arbitrariness and possible bias that the historical documents embody should be acknowledged.

As specified above, research around historical language variation has mainly been based on existing written documents produced mainly by a power elite, i.e. men who could afford classical education (Auer et al. 2014: 10). When working with a corpus containing such texts, it is essential to emphasize that the results obtained correspond only to a certain language variant – the standard language of the privileged male layer of society – and that it does not necessarily reflect the spoken language of the time nor the language of society as a whole: '[t]he linguistic forms in such documents are often distinct from the vernacular of the writers, and instead reflect efforts to capture a normative dialect that never was any speaker's native language' (Labov 1994: 11). Thus, certain types of language change may be visible in the language of the educated power elite while these changes might not be a reality in the language of the members belonging to the lower social classes. Choosing the right documents to obtain a well-balanced corpus, respecting authenticity and representativeness, is therefore crucial as 'whatever is missing from the corpus, will also be missing from all subsequent analyses, and overrepresentation as well as underrepresentation will be hard to adjust afterwards' (Piotrowski 2019b: 13).

The language of the labouring poor, representing the majority of the English population in the 18th century, 'are to date rarely represented in diachronic corpora' (Auer et al. 2014: 10). The relief letters that form the LALP corpus are a good example of this: they do not seem to have attracted much attention and might have been somewhat overlooked so far in linguistic study. Naturally, the letters do not correspond to all pauper letters that were written and only represent the documents that have survived and have found their way into for example the LALP corpus, but this does not mean that they are not valuable as linguistic evidence. These documents have rarely been studied which means that the language of paupers has been mainly ignored in language history.

These documents are nevertheless important traces of what vernacular language could have been. The phonetic spelling found in the letters might in time give a good insight into what the spoken language of the lower social classes was like compared to the language of the elite that has already been widely studied.

To be able to create a corpus as a model that respects the main criteria for corpus design, there is also another important aspect that is sometimes neglected by researchers that do not stem from the Humanities: researchers compiling the corpus need to be able to read and understand the texts that enter the corpus, even if these are in Old English, Middle English, Early or Late Modern English. As Rissanen (2018) underlines, the computers used for quantitative analyses will not be able to understand the texts nor interpret them as they have not been designed to do so:

> the computer only stores sets of data and organizes and lists them rapidly and efficiently. In the analysis, synthesis and conclusions, the machine does not replace the human brain. We will be able to ask the right questions, draw inferences and explain the phenomena revealed by our data only if we develop a good overall mastery of the ancient language form we are studying. (Rissanen 2018: 9–10)

If researchers do not know what texts the corpus is composed of for not having read nor understood them, the risk is great that errors or biases might be incorporated in the interpretations. This is why designing a corpus mindfully is 'of utmost importance for the validity of any results derived from the corpus' (Piotrowski 2019b: 13). In the Humanities, many elements and objects from the past need interpretation as they vehicle information that can be imprecise, questionable, or even totally contradictory, and 'such deficiencies then affect every further step of data processing' (Windhager et al. 2019: sect. 2). If corpus design is not given much thought, the outcome will most probably turn out to be biased and/or unreliable (see 2.1.5). Corpus analysis can also prove to be a challenge depending on for example social factors and ideologies to be taken into account, even though these are not specifically mentioned anywhere in the corpus material. The texts of a corpus may also contain language variation, phonetic writing and non-standardised spelling that obscure immediate understanding and would need clarification. The LALP project is an excellent example of careful corpus design: only pauper letters enter the corpus, all documents are read and transcribed by several team members, and spelling nor-

malisation as well as the metadata are carefully cross-checked. In case there is a document that does not correspond to the criteria of a pauper letter, it is excluded from the corpus.

In the light of what precedes, it can thus be argued that a corpus can function as a model of a historical linguistic phenomenon as long as the corpus design is efficient, the corpus elements are machine readable, the criteria of authenticity, balancedness and representativeness are respected and that it is possible to make reliable data interpretations in connection with other historical factors not necessarily present in the corpus. If the elements of the corpus are unambiguous, explicit and coherent (for example by transcription and/or language standardisation), they will entirely correspond to the requirements of the machine-readable formal model. There are, however, certain issues tied to corpora and concern the so-called 'bad data' problem. When elaborating hypotheses, theories, and when drawing conclusions, these issues are to be kept in mind.

Before discussing the questions of uncertainty, bias and the 'bad data' problem that are important aspects of the corpus and will be dealt with in subsequent subsections (see 2.1.5 and 2.1.6), I will in the next subsection describe how to make a historical corpus machine readable.

2.1.4 Making a historical corpus machine readable

There are many challenges when transforming a hand-written historical text corpus into a machine-readable format. As already mentioned, problems such as spelling variation and oral writing arise, and it is necessary to solve such issues before it becomes possible to analyse a corpus through computational methods. In this subsection, I will first briefly present the background to standard English and how that relates to the LALP pauper letters. Then I will continue with a description of the necessary processes in order to obtain a machine-readable corpus, i.e. the processes of transcription, spelling normalisation, TEI for historical texts and annotation. I will also briefly mention automatic modernisation approaches even though these may not yet be totally functional for historical texts.

The language in historical documents can prove to be a complex question to deal with in an electronic linguistic corpus that should be searchable. Language undergoes constant evolution, and the further back in time we go, the more linguistic differences we find. In England, standardisation of the language started as early as the 15th cen-

tury (Görlach 1999: 473), probably with the introduction of the printing press by William Caxton in 1476. In order to be able to print written texts and make them readable to all, it was important to choose a standard orthography. Since the King or Queen and the court were based in London, it was not surprising that 'one of the varieties spoken and written in London [became] the model for what evolved from being a synecdochic dialect to a fully elaborated standard' (McColl Millar 2012: 80). The chosen standard language thus emerged from members of the upper social classes that were well-bred, who used 'scholarly discourse' (Beal 2004: 91) and who were 'associated with the political, commercial, and academic centre of London' (Auer 2012: 940). The members of the lower social classes and the poor, however, did not know how to write standard language and probably did not participate greatly in the process of language standardisation. Gradually, the Queen's English, this linguistic variant from the South supposed to be grammatically correct and apparently better suited for formal use (Görlach 1999: 463), spread to the whole of England. In the 18th century, codification, prescriptivism and standardisation of both written and spoken English was much debated and set into practice through dictionaries, language guides, grammars and public debates (Beal 2004; Auer 2012). As for the spoken form of English, Received Pronunciation (RP) was designated as a 'neutral' standard and has ever since been promoted and taught as the correct spoken English (Crowley 1997; Trudgill 1999; Crowley 2003; Milroy & Milroy 2012).

As the pauper letters from the LALP corpus were written by people from the 'occupational classes such as shopkeepers, lesser landholders, master craftsmen, artisans, soldiers, clerks and other business assistants, labourers, servants, pedlars, publicans and paupers' (Fairman 2011: 39) with limited schooling, the letters present frequent spelling variation and phonetic writing, mirroring non-standard language variants. Since paupers did not have access to the standardised written forms of certain words they had to use, many letters exhibit phonetic writing, showing that people often wrote as they spoke. Paupers might also have guessed through analogy how words could be written, as '[a]ural writers tend to spell the sounds in their heads phonemically' (Fairman 2007: 177). This proves to be a major challenge when these documents need to be made machine readable, as the language does not correspond to any linguistic standard and thus requires careful transcriptions and spelling standardisation.

I will now briefly present some of the basic steps that are necessary to make a corpus, such as the LALP corpus, machine readable: transcription, spelling normalisation, TEI for historical texts and annotation. Finally, the new technology of automatic modernisation approaches will be succinctly presented.

Transcription

Working on hand-written historical documents implies that the corpus designers have read and are familiar with the texts that enter the corpus and that the diverse linguistic issues such as phonetic spelling, capital/small letters and self-corrections have been approached. The next step is to transcribe the hand-written documents to obtain a plain text version. Interpreting and manually transposing the hand-written text to a machine-readable format, for example .txt or .docx, is time-consuming and requires specific care to avoid as many errors as possible. Generally, the transcriptions are generated manually, since OCR (Optical Character Recognition) methods work best with printed texts and not hand-written ones. When the historical manuscripts have been transformed to a machine-readable form, it is then possible to start with computer-aided spelling normalisation, combined with additional manual checks. Transcription of historical documents is no minor task, especially when the documents are not written in standard English, which is the case with the pauper letters in the LALP corpus. Therefore, a considerable amount of time has to be dedicated to the transcription process, as several people need to transcribe each document, check and compare each transcription in order to minimise any mistakes and also to produce a corpus which will be functional and reliable. The letter sent by Moses Tyson on the 4th December 1828 (Fig. 2.1) is a good example of the challenges that scholars are facing when transcribing original hand-written manuscripts from the late 18th and early 19th centuries. The transcription of this pauper letter is the following (©LALP/Anita Auer):

Millom, BPR10/05/2

Whithaven December the 4 - - - 1828
Mr hartleey Sir I am Sorey that I have to Right
a Gain But hard Need Maks Me Do it for our
Money is Dun as it will be 2 Months Since we Gott
it be for I Gett it and it only Leaves hus onley 1=S= =2=d=

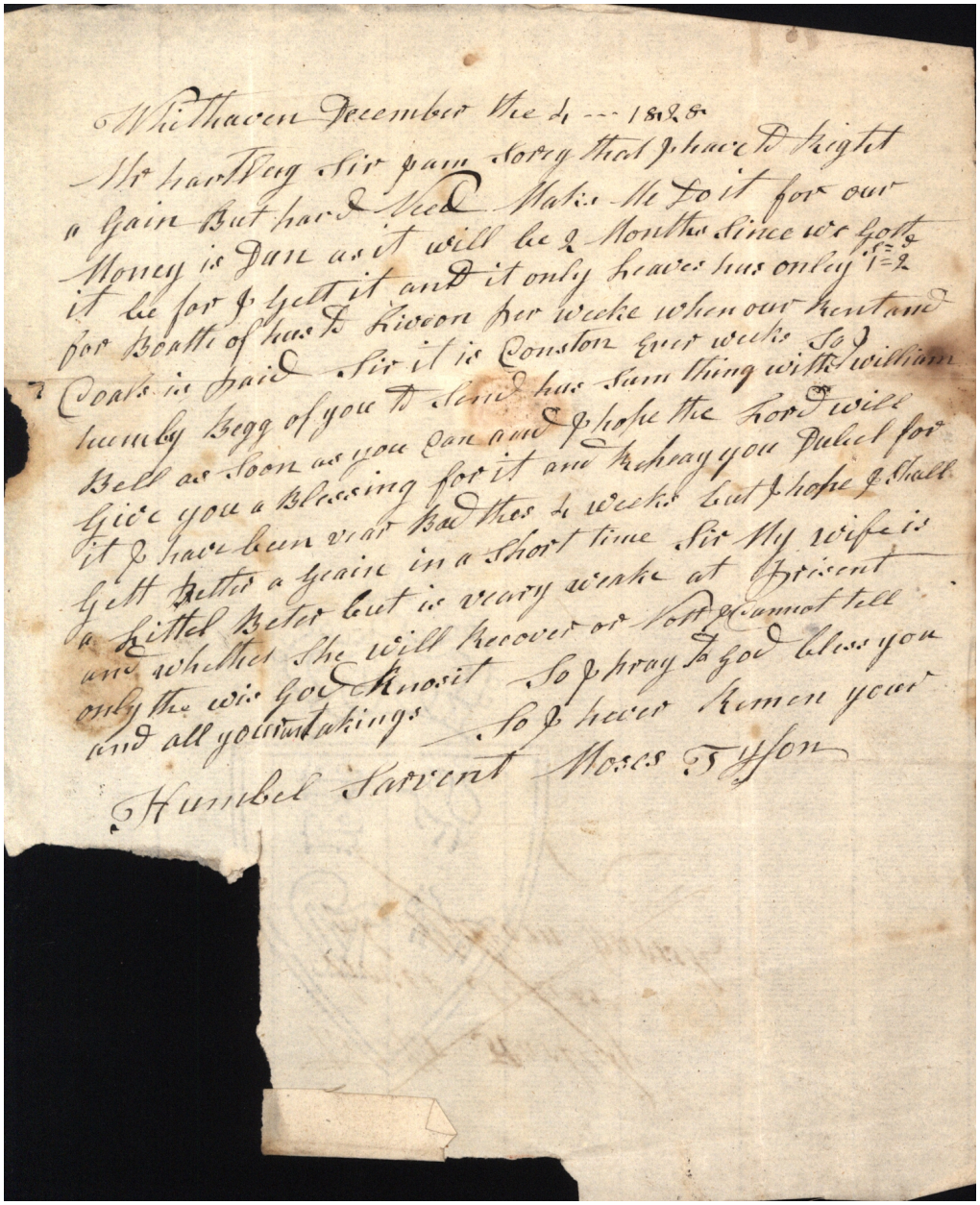

Figure 2.1: *Facsimile of a petition letter, reused with permission of the Cumbria Archive Centre, Barrow-in-Furness (Ref: BPR10O52).*

for Boath of hus to Live on per weeke when our Rent and
Coals is paid Sir it is Conston Ever weeks So I
humby Begg of you to Send hus Sum thing with william
Bell as Soon as you Can and I hope the Lord will
Give you a Blessing for it and Repeay you Dubel for
it I have been vear Bad thes 4 weeks but I hope I Shall
Gett Better a Geain in a Short time Sir My wife is
a Littel Beter but is veary weake at prisent
and whether She will Recover or Nott I Cannot tell
only the wis God Knosit So I pray to God bless you
and all your [^un INSERTED ^]takings So I hever Remen your

Humbel Sarvent Moses Tyſon

Moses Tyson's letters shows random capitalisation, phonetic writing, hypercorrection, lack of punctuation as well as a preference for Anglo-Saxon monosyllabic words. The hand-writing, in this case quite refined, and phonetic spelling are important factors when deciphering a pauper letter. Paupers were generally not taught how to write standard language and only had access to drawing graphs (or mechanical writing), i.e. they were capable of writing each letter separately (Auer & Fairman 2013: 195). Tyson clearly drew the graphs carefully in spite of the so-called mechanical writing which is frequent in pauper letters. At times, the scholar needs a great handful of imagination and deduction abilities to figure out what is written and what was intended because of the phonetic spelling. The lack of punctuation is also a factor that might render the transcription laborious. The characteristics of pauper letters will be dealt with later on in this section (see 2.2.2).

Spelling normalisation

In the texts of a historical corpus such as the LALP relief letters, the spelling may vary considerably. At the end of the 18th century, standard English was widely debated but not yet fully in place, and the lack of general education for all, regardless of social status and fortune, was probably an important obstacle in its adoption. When one reads the

relief letters, the first striking feature to be noticed is the impressive amount of spelling variation, even in one single letter. From the 16th century onward, the spelling in printed texts also varied significantly in spite of a growing standardisation of the English language, but 'it would seem that the variety of spelling we find in printed works in the Early Modern period, which is already spectacular, would pale beside the variety used in handwritten papers' (Craig & Whipp 2010: 39). So it is not surprising that relief letters from the Late Modern Period exhibit the same extensive variation, as the pauper applicants had not had access to schooling. Therefore, if texts presenting important oral writing and spelling variation are to be analysed with quantitative methods, it is essential to normalise the language before any further research can be undertaken, as Marquilhas and Hendrickx highlight:

> the lack of normalisation for spelling creates a problem when letters are seen as a target for corpus linguistics operations: morphologic annotation, parsing, semantic annotation, concordancing, word lists, and keywords. Such level of processing demands for a *corpus* in standard spelling, a resource also invaluable for historians focusing on the discursive features that manifest themselves through keywords and semantic fields present in the corpus. (Marquilhas & Hendrickx 2014: 67)

However, the amount of computer programs that allow for spelling normalisation of historical texts presenting spelling variation are few. Today, the main tool, VARD2 or the VARiant Detector, was developed by Alistair Baron and Paul Rayson (Baron & Rayson 2008; Archer et al. 2015) for Early Modern English. This program has also been adapted to other languages than English. VARD2 allows spelling normalisation in an efficient way, executing spelling normalisation on enormous quantities of texts in a minimum of time, in spite of certain initial manual tasks, necessary to build up the variant dictionary. Even so, 'such normalisation needs to be handled sensitively: so that, for example, we can maintain – within the text – the original spelling of those forms which convey important morphosyntactic or orthographic information' (Archer et al. 2015: 6). If words are manipulated in such a way that their original form completely disappears, the risk is that information about the text is lost. It is also important to remain aware of the fact that computer programs are not infallible, and that they need to be trained on data, either through manual or automatic standardisation, evaluated and improved to enhance training capability (Baron & Rayson 2009: 9–13).

In Statistical Machine Translation (SMT), the spelling normalisation process is considered to be a process of translation. Instead of focusing on words and phrases as in traditional translation, the SMT method is based on the idea 'that phrases are modeled as character sequences instead of word sequences' (Pettersson, Megyesi & Tiedemann 2013: 56). Even a small training corpus is sufficient for making SMT efficient. In their research, Pettersson, Megyesi & Tiedemann (2013) used word pairs combining one word with modern spelling and the other one with historical spelling. This method is not only valid for English but for other languages that have been standardised. Furthermore, Marquilhas & Hendrickx (2014: 70–74) describe how the VARD2 program was improved for the Portuguese language by the creation of DICER (Discovery and Investigation of Character Edit Rules), 'a statistical tool that creates a list of edit rules on the basis of a corpus labelled with spelling variants and their modern counterparts' (Marquilhas & Hendrickx 2014: 70). Even though an important amount of work was done thanks to the computer, there was still much manual work involved in the improvement process, for example manual transcriptions from historical language to modern language and manual checks of certain recurrent spelling variants, as the DICER project team 'had already noticed that some variants were not mapped to a modern word form but to another, more frequent archaic word form' (Marquilhas & Hendrickx 2014: 75). Although computers are incredibly efficient when dealing with automated tasks, there are still errors that need to be corrected manually by human researchers. This manual part of the work is, however, time-consuming and should not be underestimated.

VARD2 can operate both on large and small amounts of texts, which is an advantage when the use of spelling normalisation is necessary for small corpora. Spelling normalisation computer programs represent both advantages and drawbacks, and for any research team and research project, it is nevertheless crucial to remember that

> [d]ealing with variant spelling is a major challenge for automated methods, and while they do make it possible to count words in old-spelling texts on a large scale, we have to accept that a percentage of error is built in, far beyond the occasional slip and the odd piece of guesswork in a text worked over case by case by an expert editor. (Craig & Whipp 2010: 49)

In spite of these cautions, any researcher that has used computational tools such as VARD2 for spelling normalisation can certify that it is immensely time-saving.

At this stage, however, I would like to stress the importance of the mastery, or at least a good working knowledge, of computational techniques when using them in research. At first sight, VARD2 can seem user-friendly and relatively easy to learn. And yet, it is of great importance to acquire solid proficiency of the program as the program menu is vast and demands some time investment to be fully integrated.

As already discussed, it is necessary to work on documents with standardised spelling if computational techniques are to be used for quantitative analyses. Therefore, when researching historical documents, turning these into electronic machine-readable standardised texts will allow texts 'to be encoded and annotated for storage and further processing' (Piotrowski 2012: 53) in view of future use and research. The encoding can be performed with different tools, but in the case of historical sociolinguistics, there is one format that is especially suitable: TEI for historical texts.

TEI for historical texts and annotation

Encoding and annotating texts is of great importance for corpus linguistics research. To make historical texts exploitable computationally, encoding methods such as TEI – Text Encoding Initiative – can be used. The Text Encoding initiative is a project that gives scholars and researchers in the Humanities guidelines as to how to encode and manage the digital versions of all types of documents (Burnard 2014).

The Text Encoding Initiative was created in 1987 by the Text Encoding Initiative Consortium 'in order to develop vendor-independent standards for encoding Digital Humanities data, in particular text' (Piotrowski 2012: 60). The TEI Consortium has developed high-quality guidelines[6] for researchers in the Humanities to follow when encoding literary and linguistic texts. TEI is an open-source project and uses the meta-language XML (Extensible Markup Language) to create electronic forms of written texts that allow the texts to be stored and processed. The guidelines contain a detailed introduction to XML that is available to users for free. Historical texts can thus be transformed into TEI XML documents that offer more opportunities for qualitative as well as quantitative research than before digitisation.

[6]https://tei-c.org/release/doc/tei-p5-doc/en/html/index.html

The advantages of TEI XML are multiple, but it is well worth mentioning three of them:

> The first is that TEI XML focuses on the meaning of text, rather than on its appearance. The second is that TEI XML is independent of any particular software environment. The third is that TEI XML was designed by and for the scholarly research community, which is also responsible for its ongoing development (Burnard 2014: 7)

As technology advances very rapidly, it is essential to encode texts in such a language that will be machine readable even after 20 or 30 years from now. Today, compatibility between different systems such as Mac, Windows and Linux cannot be taken for granted, and even opening a recent document on different word processors such as Word or OpenOffice might prove to be complicated or, in the worst cases, impossible. Therefore, using the open-source TEI XML metalanguage will guarantee a continuity in legibility.

The TEI Guidelines website[7] available in several languages provides all the necessary information concerning how to transform documents using TEI methods. Learning the code and using it spontaneously does require time and efforts, however, as the XML metalanguage is not exactly the mainstream of computer languages taught in pedagogical institutions today and prior knowledge of this type of coding is generally low. For researchers in the Humanities, learning to master this kind of computer language and methods can seem intimidating, even though the advantages of doing so might prove significant. Yet, it is true that technology advances swiftly, and one computer program might be replaced by another after a short period of time, which discourages more than one to put time and effort into becoming even more computer literate. But even though the TEI XML language has developed over decades, it still provides a good long-term growing ground for historical document encoding, and it is well worth the initial learning effort.

Moreover, prepared TEI documents allow for annotation of texts, which can prove to be important when additional data must be preserved in relation to the text. Annotation is 'the process of enriching a collection of text by adding linguistic and inter-

[7] https://tei-c.org/release/doc/tei-p5-doc/en/html/index.html

pretative information to it' (Alvarez-Mellado et al. 2021: 527). Explanations of linguistic elements and metadata about the document itself and/or of specific words can thus be encoded to become accessible to each user and shed new light on factors closely linked to the original texts. In historical sociolinguistics, annotating a document can be extremely useful, as this allows for the preservation of original word forms that would otherwise disappear in the spelling normalisation process. Today, there are specific tools such as the TEI Publisher[8] that allow scholars that are not programmers to annotate the original documents without any extensive prior knowledge of XML. Annotation can preserve valuable information that would otherwise be lost, and if this information could not be stored, research teams would have to provide extensive new research each time the document was under inspection.

TEI and spelling normalisation with VARD2 are excellent tools to manage different steps of the process of corpus creation. Today, new methods are being developed to minimise the manual work as much as possible. Automatic modernisation with AI is such as method and seems to be a promising tool for future work, even though there is still room for improvement.

Automatic modernisation approaches

As described above, transcribing and normalising historical texts require much time and efforts by an important number of humans when dealing with an extensive corpus. It is possible to use a spelling normalisation program such as VARD2 to transform non-standard spelling into contemporary standard language, which alleviates the process of normalisation. But even working with VARD2 is a time-consuming task, although a part of the translation work is being done by the machine. There is still a certain amount of manual manipulations to be executed, even with a computer program specifically designed for such a purpose as spelling normalisation. Because historical documents have become of great interest to researchers from different fields over the last three decades, efforts are being made to minimise these manual and time-consuming steps as far as possible through automatic modernisation approaches. Such methods are for example Character-level Statistical Machine Translation (CSMT), Statistical Machine Translation (SMT) and Neural Machine Translation (NMT). These approaches are developing rapidly but

[8]https://teipublisher.com/exist/apps/tei-publisher/doc/documentation.xml?id=introduction

are however not yet totally functional and still need extensive testing and refinement to produce trustworthy results.

Two questions that are frequently raised in connection with modern spelling normalisation is first to what degree the original text loses its substance in the transformation, especially with automated transformation, and then whether automatic modernisation approaches are 'sensitive' to human language, i.e. if the computer programs are capable of interpreting texts in the same way and putting meaning into the transformations as humans do. Of course, some important data are irremediably lost when a 18th-century pauper letter is transcribed and the spelling is normalised. Domingo & Casacuberta (2020: 1) argue that modernising language of historical documents can be considered as a questionable process since the language of the original document is considerably changed through substantial modifications in language rhythm and rhyme. This loss of what could be considered as the core of the historical document is nevertheless counterbalanced by easier access to a historic past that would otherwise be rendered extremely difficult because of language forms unknown to us today. The modernisation of historic texts is nevertheless important in order to have access to a literary and cultural heritage from the past, as well as for linguistic research. This process as such is not new, as '[m]odernization has been manually applied to literature for centuries' (Domingo & Casacuberta 2020: 2). *The Bible* and classical literature written in Latin and Greek are basic examples of this. Literary works by Chaucer, Cervantes and Shakespeare have also enjoyed extensive translation, and all these texts are still under constant investigation (Domingo & Casacuberta 2020).

Modernising the language of historical documents and digitising the texts allow for new research, for example through corpus linguistics, which means that researchers may detect unfamiliar patterns that were not distinguishable before through qualitative research only, as Domingo & Casacuberta (2020: 2) point out. In their research on machine translation, the use of back-translation for their project as well as making a user study, Domingo & Casacuberta evaluated the performance of Neural Machine Translation (NMT), based on whole sentences, compared to Statistical Machine Translation (SMT), based on words and phrases. In their research, they included both automatic and human evaluation. Interestingly enough, 'the human evaluators slightly preferred SMT over NMT' even though 'the automatic evaluation [...] did not show any significant dif-

ferences between the SMT and NMT approaches' (Domingo & Casacuberta 2020: 6). In their conclusion of this project, these two researchers state that users had a preference for the SMT approach: '[w]hen comparing the SMT and NMT approaches, the NMT approach made a bigger number of errors and the user chose its modernized versions as the best option fewer times than with the SMT approach' (Domingo & Casacuberta 2020: 8). These results are interesting, but the reasons for this are apparently not known.

As already mentioned above, automating processes include the training of computers to act in a certain way by using training and test sets, rendering research more complex. As Mäkelä et al. (2020: 92) point out, 'it is not an exception but the norm that projects in the digital humanities and social sciences have to deal with [...] unprecedented levels of data complexity' and continue to underline the importance of openly sharing data around these new technologies in order to correct errors and maximise the outcome: 'the overall research process can become much more efficient if people share their data-related discoveries, documentation and clean-up pipelines openly, making them available for reuse and further modification, for instance based on open licensing' (Mäkelä et al. 2020: 93). When developing new computational tools such as NMT, it would be essential to imply scholars from all disciplines concerned with the topic, such as linguists and translators as well as data scientists and computational linguists. Digital Humanities allow for such multidisciplinarity, and I am personally convinced that it is only through interdisciplinary work and collaboration between scholars and researchers from different specialist fields that technological improvements will become fully operational. This will probably also allow bias and stereotypes in automated processes to be kept at bay.

The questions around automatic modernisation approaches lead us to other important issues linked to the corpus, namely uncertainty and bias that I will briefly present in the following subsection.

2.1.5 Dealing with uncertainty and bias

I will now discuss two factors that need to be taken into consideration in corpus design and when using a corpus for research: uncertainty and bias. These two concepts represent two separate phenomena but both may influence on the creation of a corpus as well as on the results of corpus analyses.

Uncertainty

The word 'uncertainty' is defined, according to the *Oxford English Dictionary*, as '[t]he state of not being definitely known or perfectly clear; doubtfulness or vagueness' (*OED online* 2021: def. 2a). When researching language in historical documents, uncertain linguistic elements and miscellaneous information in connection with language are often encountered with. There might be missing dates and spelling variation in names as well as other uncertain data. Uncertainty is frequently related to birth dates, personal data, truncated place names, undetermined graphs, spelling variation giving rise to unclear meaning, author authenticity or contradictory evidence, such as variations in author names (Binder et al. 2014: 96). Uncertainty is not only a lack of information but 'can have different causes, take different forms, and is related to other phenomena such as imprecision, vagueness, and ambiguity; it may also involve issues of belief and trust' (Piotrowski 2019a: Introduction, para. 4). Some data, or rather predictions, are always uncertain, such as the weather forecast as well as in explicit uncertainty, such as the phrase 'He is between 50 and 60' (see Piotrowski 2019a). This kind of uncertainty can usually be confirmed: it is possible to wait a few days to confirm or refute the weather prognosis, or we can ask our interlocutor for more information about somebody's age. At times, language can be imprecise or utterances can be misunderstood as '[n]atural language is full of phenomena of ambiguity and uncertainty' (Andresen, Vauth & Zinsmeister 2020: 48).

In spite of such ambiguities, language is not an aimless act of communication. Humans use language with a specific result in mind, expecting certain outcomes or consequences to their utterances: '[l]anguage is not random' as we 'speak or write with purposes' (Kilgarriff 2005: 264). These purposes are not, however, always clear and require some interpretation from people participating in the linguistic exchange. The meaning of a specific utterance might not be evident and can thus create ambiguity and uncertainty. If someone says 'It is chilly in here', does the person simply mean that s/he feels cold, or could it be a polite way to ask the host to close the window? Or on the contrary, perhaps the speaker wishes to borrow the fabulous blue cardigan thrown on the sofa? The purpose of the utterance is thus ambiguous and it would therefore be important to have a context in order to be able to make the right interpretation and define the meaning of what is being said or written. In historical documents, the context may be relatively

vague, and there is only little evidence that can bring clarity into such uncertainty. As 'we do not yet have a standard procedure for integrating ambiguities in formal models—or even for identifying ambiguities in the first place' (Andresen, Vauth & Zinsmeister 2020: 48), it could prove to be important to create models of uncertainty according to the needs of the research question if uncertainty and/or ambiguity is a major part of the research project.

There are cases when reliable information is simply inaccessible. This can be the case with missing data concerning an object or historical event, contested or imprecise data, such as several dates of birth (Windhager et al. 2019: sect. 5). Uncertainty of this kind 'can never fully be resolved, as we will never have perfect knowledge when we are dealing with the real world' (Piotrowski 2019a: Introduction, para. 4). Since the 'real' world is imperfect and it is not possible for humans to have foolproof evidence in every situation where tangible proof would be required, it could be argued that uncertainty is certain, which sounds like a paradox. Surprisingly enough, researchers in the Humanities are constantly being confronted with uncertainty, especially when it concerns historical documents and/or historical objects that are fragmentary. This entails that if uncertainty needs to be formally modelled, it is necessary to determine two basic aims: 'to make uncertainty explicit' and 'to allow reasoning under and about uncertainty' (Piotrowski 2019a: 4. Modeling uncertainty, para. 1). But as Piotrowski argues, it is impossible for computers to deal with for example imprecise information or unstructured text fields such as 'late 13th century' or 'sometime between 1291 and 1295' as the computer program requires a precise date, for example '1294' (Piotrowski 2019a: 4. Modeling uncertainty, para. 4). Uncertainty can thus represent an obstacle to the conversion of historical data to a machine-readable format if there are no text fields that accept uncertainty. Before it becomes possible to model uncertainty, it would be necessary to precisely define what uncertainty is and to determine what typical characteristics uncertainty shows (Piotrowski 2019a). It could be argued that uncertainty in the Humanities is different from the types of uncertainty to be found in for example hard sciences. Piotrowski (2019a) gives a few suggestions as to what uncertainty in the Humanities is: it is 'rather qualitative than quantitative, or at least hard to quantify', data that 'often concerns singular, non-repeatable events', and 'phenomena similar to those known as ***selectively reported data*** and ***missing data*** in statistics may be more frequent' (Piotrowski 2019a: 5. Conclusion, para.5).

This is why theoretical Digital Humanities would be needed in order to develop this issue further (Piotrowski 2019a: 5. Conclusion, para. 8).

Considering this, in what way is it possible today to deal efficiently with a lack of data and uncertain data in order to make attempts at modelling uncertainty? Uncertainty has been much discussed in computer science, but in Digital Humanities, this tendency still seems modest (Piotrowski 2019a: 4.2 Uncertainty in Digital Humanities, para. 3). Uncertainty is dealt with when it arises in specific projects, such as the digitisation of artefacts (Tarte 2011) or the construction of a georeferenced online bibliography of Holocaust and camp literature (1933-1949), project carried out by Binder et al. (2014). This project called GeoBib was interdisciplinary, involving scholars from the fields of history, literature, geography and computer science. The team succeeded in modelling uncertainty in various ways, depending on the nature of the data (Binder et al. 2014: 96):

- by inventing a MediaWiki System that allows for vague and uncertain information
- by creating separate XML files for inconsistent information, such as differences in editions
- by providing the @cert TEI attribute with annotations that can easily be read by human readers, as well as <note> elements that explain the nature of the uncertainty: '[t]he combined use of both elements allows conveying uncertainty to the human user while keeping it encoded in a machine-readable (or machine-traceable) way' (Binder et al. 2014: 96)

As indicated, these examples are tailor-made solutions for a specific project (see also Marquilhas & Hendrickx 2014). Piotrowski argues that the Humanities still lack 'a *systematic account of uncertainty in humanities research*, which would aim to document causes for uncertainty, as well as its behavior' (Piotrowski 2019a: 5. Conclusion, para. 2). As uncertainty 'is an omnipresent property of information and knowledge in a given field', it would seem relevant to gain a deeper insight into modelling uncertainty and 'to establish a more nuanced understanding of this concept' (Windhager et al. 2019: sect. 5, para. 1). It could be argued, though, that Humanities scholars as well as visitors of museums and other members of the general public are familiar with uncertainty,

which would minimise the need for modelling it, since '[m]ost of non-digital humanities research is plausibly dedicated to actually coping with this challenge' (Windhager et al. 2019: sect. 5, para. 4). For the general public, conveying uncertainty can be done through for example texts and other language-based elements.

If uncertainty is to be modelled, it would first of all seem essential to understand what kind of uncertainty there is, for whom the model is intended, if it is imperative and useful to make a model of the uncertainty and how it could best be modelled. To model uncertainty, for example for cultural collection visualisation, it is therefore indispensable to know in what form the uncertainty appears, for whom the model is important (experts or casual users) and what there is uncertainty of (Windhager et al. 2019: sect. 3-5). Windhager et al. suggest two solutions for modelling uncertainty: on one hand, it would be reasonable to model uncertainty because 'trust for DH experts is enabled and deepened by transparent and truthful system designs' and on the other, to take into consideration 'that trust for casual users does not emerge from a detailed structural and procedural understanding, but rather from an assembly of contextual cues' (Windhager et al. 2019: sect. 6, para. 5). The target audience is therefore of importance when considering modelling uncertainty as well as the various expectations stemming from both professionals and laymen.

It is therefore not only spelling normalisation that can benefit from digital advances and automation tools, but also the creation of uncertainty models. The importance of probabilistic programming as a malleable tool when dealing with uncertainty and bias is put forward by Lahti, Mäkelä & Tolonen (2020: 281). In his livebook available on the Web, Pfeffer (2016: chapter 1, 1.1) also argues that probabilistic reasoning 'combines our knowledge of a situation with the laws of probability to determine those unobserved factors that are critical to the decision' and then goes on to explain what probabilistic programming is. Lahti, Mäkelä & Tolonen (2020) also discuss the use of probabilistic programming as follows:

> Probabilistic programming can be used to build explicit models and compare evidence between alternative hypotheses on the data generating processes. [...] [T]his can help to bridge the gap between qualitative and quantitative interpretations, and provides promising tools for hypothesis-driven, data-intensive research in computational humanities. (Lahti, Mäkelä & Tolonen 2020: 281)

Probabilistic programming is, however, not readily available to scholars in the Humanities because of its complexity, reason why interdisciplinary projects are of the highest importance, ensuring that each scholar and expert will work with what s/he masters. Lahti, Mäkelä & Tolonen (2020: 287) argue that 'effective construction, use and interpretation of probabilistic models requires a robust understanding of modern statistics as well as statistical programming'. In my opinion, it is therefore unrealistic to expect an art historian or a linguist to transform into a data scientist or programmer when encountering uncertainty in a research project. Not only is it important for researchers to establish fruitful collaborations that will benefit from each researcher's expertise, but as mentioned above, it is also essential to know who the final user is and how s/he can best benefit from uncertainty models. Binder et al. (2014: 97) highlight that '[t]he encoded uncertainty has to be communicated effectively to the human user', which means that formal models of uncertainty and the information they vehicle should be profitable not only to professionals but also to the general public whenever the topic is of interest to them.

Modelling uncertainty seems therefore possible after having pondered what is to be modelled, for whom the model is useful and if the model can contribute to trust-building or efficient information transfer. In spite of the difficulties connected to the modelling of uncertainty, it is nevertheless feasible to create models for example through probabilistic programming, not only in DH but also in other fields such as product marketing[9].

Bias

Uncertainty is not the only element that can have an influence on the corpus design and on the research results based on the corpus. A corpus can also contain biases stemming from human behaviour and existing human attitudes and performance that do not reflect impartiality. A corpus can contain biases for several reasons, for example because of uncertainty, or because 'the corpus will be constructed in a way that can only serve to confirm the analyst's pre-existing expectations' (Gries & Newman 2018: 257).

[9] see Pfeffer 2016

What then is bias? According to the *Oxford English Dictionary*, the word 'bias' has many definitions, but the most pertinent one in this context is 'any preference or attitude that affects outlook or behaviour, esp. by inhibiting impartial consideration or judgement' (*OED online* 2021: def. 1: 3c). The element of impartiality is important: In order to work correctly and yield pertinent results, a corpus should be designed as impartial and objective, respecting the criteria discussed previously, leaving aside any preconceptions, stereotypes, uncertain data (if the uncertainty is not modelled or taken into consideration) or other discriminating features. Therefore, continuing with the example from the previous subsection on uncertainty, what would happen if the date '1294' were introduced into the corpus instead of keeping the uncertain information 'late 13th century', with the aim of making the corpus machine readable and easier to search? This date could be considered as faulty data and thus create bias in the final results of the analysis, as certain conclusions might be drawn according to the possible preference of the researchers for that exact date.

Bias can also arise because the surviving historical documents or historical objects 'can not be considered to be a random, representative sample of what was, as they reflect the interests of collectors' (Mäkelä et al. 2020: 87). Many historical objects, such as paintings and books, have survived thanks to private collections and, later on, libraries. Some objects might have been preserved in accordance with personal or arbitrary preferences while others might have been discarded or intentionally destroyed for political or monetary reasons. Using library catalogues to create formal models also vehicle an increased risk of bias, considering 'the long time span under which the cataloguing process has been executed' (Mäkelä et al. 2020: 87). When researching historical documents, linguists and historians need to rely on such published books and data that are available in the catalogues, on any remaining microfilms and perhaps also on 'a long process of manual curation with different actors and according to different standards' (Mäkelä et al. 2020: 81). All these elements might convey uncertain, limited or biased data, which could be an issue for corpus design and analysis interpretation. Therefore, it is important to publicly state and explain that such biases can interfere with the final results of research based on corpora presenting possible biases. For example, when researching the language in historical documents, it is important to take into account that language is tied

to the historical context and social conditions of the time period. Such aspects cannot be ignored and could bring about biases if they are.

Another important and so far little investigated research field is machine bias. Today, through Artificial Intelligence (AI), machines are taught to behave in a certain way, using either supervised or unsupervised learning[10]. Training and test sets are constituted for their specific task (training of machines), and these will subsequently allow algorithms to perform certain automatic tasks on corpora or sets of data (see Sun et al. 2019). Nevertheless, if biases are introduced to the training and test sets, these will automatically be reproduced later on in the results obtained[11]. And although scholars and researchers should try to avoid bias as much as possible, it is apparently no fatality – bias 'can, however, be potentially detected and treated through explicit formal analysis' (Lahti, Mäkelä & Tolonen 2020: 280). If the risk of bias or the bias itself is identified at an early stage when designing a corpus, it can be corrected for example through probabilistic programming, which can lead to a reduction of the cleavage between qualitative and quantitative research.

Bias in Natural Language Processing (NLP) and Machine Learning (ML) has only recently become a field of interest to researchers, and especially problems concerning gender, racial or facial recognition biases have been highlighted as possible pitfalls for future work with models: 'Although NLP models have shown success in modeling various applications, they propagate and may even amplify gender bias found in text corpora' (Sun et al. 2019: 1630). When working on a project on machine translation with Google Translate, Prates et al. discovered that 'statistical translation tools such as Google Translate can exhibit gender biases and a strong tendency toward male defaults' (Prates, Avelar & Lamb 2018: 6377). Almost immediately after the publication of this article encouraging translation engineers to be careful about the training sets, Google changed its policy in order to be able to offer the users 'a new feature presenting the user with a feminine as well as a masculine official translation', this being 'part of a broader goal of promoting fairness and reducing biases in machine learning' (Prates, Avelar & Lamb 2018: 6376). Other researchers have recently also highlighted other cases of gender bias for example in ELMo (Embeddings from Language Models) contextualised word embeddings (Zhao et

[10] https://en.wikipedia.org/wiki/Machine learning

[11] Amazon's sexist AI recruiting tool, https://bit.ly/3EK9EPb

al. 2019). These questions are extremely interesting and of highest importance, but I will, however, not treat them in more detail as they go far beyond the scope of my *mémoire*.

Technology is advancing fast, but AI and new computer programs are not always the only nor the best solutions for certain research projects. When working with linguistic corpora, using computers and digital data saves research teams much time and efforts, since machines can perform repetitive and time-consuming tasks that would otherwise be practically impossible for humans, for example quantitative and statistical analyses. But even quantitative data need interpretation, and up to today, machines are not yet capable of interpreting research results. Over more than two decades, it has become apparent that humans are (so far) better than machines at interpreting results of for example language corpora:

> I would like to emphasize that the data so readily offered by corpora are useful only if they can be correctly interpreted. The computer does not replace brainwork: the value of the results depends entirely on the scholar's competence in the language form of the period or periods s/he is studying. For this reason, corpora should never be allowed to devalue a thorough learning and knowledge of the early periods of the English language and readiness to study original texts. These skills are the only way for us to make full use of the wider vistas offered by electronic tools. (Rissanen 2000: 14)

It is clear, however, that the combination of qualitative and quantitative methods allows scholars and researchers to contemplate the material from several angles, and it may also give rise to new perspectives and research questions. But as Rissanen points out, it is important not to lose knowledge around the original texts revealing gradual linguistic development. That is why linguists and sociolinguists are indispensable keys to understanding language use and language change, as machines are not yet able to interpret results holistically nor are they capable of taking into account what could be identified as human irrationality.

Now that the concepts of uncertainty and bias have been treated, I will continue in the next subsection with the concept of the 'bad data' problem that is frequently referred to by scholars in historical sociolinguistics.

2.1.6 Dealing with the 'bad data' problem

We have already seen that historical texts and documents have not been systematically preserved, which means that the documents we are working with today have survived more by chance than by any general wish for text preservation (Hernández-Campoy & Schilling 2014: 66). This is, of course, one aspect of what is called the 'bad data' problem. The historical data we are in possession of today can be considered as random and not particularly representative of a specific aspect we would like to research, for example spoken language. Written texts do not represent spontaneous spoken language, and it is very difficult to be certain that written records actually represent spoken language of a specific time period (Labov 1994; Kytö & Walker 2003). In addition to this, there is a great deal of socio-historical information that is not available to us today, and it is often necessary for researchers in the field of historical sociolinguistics to reconstruct the prevailing socio-economic and socio-cultural factors that might have influenced on language change and variation (Labov 1972, 1994; Nevalainen 1999).

The 'bad data' problem is not an uncomplicated issue in historical sociolinguistics, as '[t]he most important disadvantage of datasets of historical documents is that they very often lack representativeness and possibly also validity, since [...] the historical record is incomplete, and written materials may or may not be reflective of the spoken language of the time period under study' (Hernández-Campoy & Schilling 2014: 66). An illustrative example are trial proceedings and witness depositions that are supposed to reflect speech but that are generally written down by scribes or clerks during or after the actual speech situations (Grund 2007). Furthermore, there may be hand-written versions of such documents as well as printed ones, and the different versions might vary according to the use and intended audience (Kytö & Walker 2003). This can of course represent a problem of 'bad data':

> We face not only such problems as the possible time distance between the original version of the source text and its later copies, the history of textual transmission, unidentified authorship, and lack of background information but also the question of faithfulness. How faithful a record of the spoken language can a trial or deposition text be, considering how difficult it is to take down speech straight from the mouths of the speakers in practice? Are these records in a sense a case of 'bad' data

> [...], or can they offer us a glimpse at early specimens of authentic speech? (Kytö & Walker 2003: 221–222)

When dealing with pauper letters soliciting out-parish relief, however, the issue of originals and copies is generally not relevant, as the letters are 'authentic accounts of how people had to make use of written vernaculars with the established societal system' (Auer et al. 2014: 26). Furthermore, there probably never were any copies made by the paupers themselves because of the cost of good quality paper (S. King 2019: 67) in contrast with the letters emanating from the elite, who regularly had access to scribes and kept copies of letters for themselves.

It is nevertheless legitimate to question whether the existing data can actually reveal important information about spoken language. As pointed out by Hernández-Campoy & Schilling (2014: 64), the documents used in historical sociolinguistics are not 'a second-best solution by inevitable necessity, but just the best solution in those areas of study for which oral records are not available'. Researchers in historical sociolinguistics often emphasize that it is important to use the data available as best as possible, even though it can be considered by some as 'bad' (Nevalainen 1999; Kytö & Walker 2003; Hernández-Campoy & Schilling 2014; Auer et al. 2015). In the case of pauper relief requests, the phonetic spelling and other characteristics (see 2.2.2) may lie closer to the speech of the writer than for example family correspondence written in standard English. Therefore, such original documents should not be dismissed only because they do not present all necessary qualities for efficient quantitative analysis. It is important, however, to acknowledge that there can be 'bad data' in historical documents and that this needs to be taken into consideration when proceeding to analyses of the research material.

The important issues of 'bad data' in pauper relief requests lie elsewhere than in originals and copies, as in trial examination records. First, it is almost impossible to know who physically wrote the relief request. Was the applicant the person who wrote the letter and signed, or did s/he have a family member, or perhaps a friend or acquaintance, who actually held the quill? How much of the writing is coming directly from the applicant and what part comes from the unnamed scribe? In the 18th century, there were still people who did not know even how to sign their names, for example in the marriage registers (Woolf & Fox 2003: 24; Laitinen & Auer 2014: 188) and even less write a relief

request. Another factor that contributes to the 'bad data' problem is that we do not know how much schooling each writer had had: was s/he trained in writing, and if so, where had s/he learnt the writing skills? Advocates might also have written letters for some paupers – if so, who were they and how did their language reflect the language skills of paupers? Many other facts are unknown to us, for example the applicant's background, her or his profession or age, the reasons s/he had had to leave the home parish, and the type of work that was available, attracting the pauper to migrate. At times, even the gender of the writer remains unknown as there is no signature, or all the letters for one single pauper are written in different hands. This kind of interference needs to be taken into account when analysing the language in speech-related texts (cf. Kytö & Walker 2003). Nevertheless, pauper letters are important testimonies of speech-like written language, as they exhibit extensive phonetic writing and 'it becomes apparent that much of language variation and changes past and present will go unnoticed if one does not acknowledge language material that is close to "vernacular" language' (Elspaß 2011: 7). Therefore, it becomes essential to deal efficiently with uncertainty, the 'bad data' problem and even the lack of reliable data for both quantitative and qualitative analyses. This can be done by using a multidisciplinary approach and by explicitly stating where pitfalls might occur. This should not, however, discourage nor stop researchers from using historical material in linguistic research.

So far in this *mémoire*, I have placed the corpus at the forefront as a general and mostly theoretical notion. Let me now continue with a specific example of what a corpus and corpus design can be by first presenting what pauper relief letters are and then by giving a brief overview of The Language of the Labouring Poor in Late Modern England project (LALP) that is currently ongoing at the Université de Lausanne. I will also discuss the question that tickles my curiosity, namely possible gendered discourse in pauper relief requests.

2.2 Pauper relief requests and the LALP corpus

In this section, I will first explain what pauper relief letters are and why they were written (2.2.1). Then I will discuss their main characteristics (2.2.2) and continue with the importance of pauper letters in linguistic research (2.2.3). A presentation of the project The Language of the Labouring Poor in Late Modern England (LALP) will then follow (2.2.4) before a discussion on the question of possible gendered discourse in pauper relief letters (2.3).

2.2.1 What are pauper relief letters?

In 1601, the Elizabethan Poor Law, or also called Old Poor Law, was passed by the English Parliament. Basically, this Parliamentary Act stipulated that 'all parishes in England [...] were statutorily required to relieve their poor' (Sokoll 2000: 21). Poor relief distribution was organised through local officials (elected overseers) of the home parish. Indoor relief could be granted after a face-to-face discussion with the applicant, while such parishioners that had moved to another region, for example to find work, had to send a written application to the parish overseer unless they could afford to travel to their home parish to apply for relief in person. Generally, a person belonged to a local community if s/he was born in the parish. One of the effects of the Old Poor Law (and especially of the Law of Settlement and Removal introduced in 1662) was to encourage the parishioner to stay in their home parish (Feldman 2003). But as unemployment rose in the late 18th and early 19th centuries and different traditional trades and professions became obsolete due to technological improvements during the Industrial Revolution, local authorities were not discontented to see the poorest part of the population move away to other parishes to find employment. Thus, the parishes could avoid signs of pauperism in their own parish by 'exporting' the poor elsewhere (Sokoll 2000).

Under the Old Poor Law and especially after the Removal Act of 1795, paupers living outside their parish of settlement were entitled to financial help, or relief, from their home parish (Laitinen & Auer 2014: 189). It was no easy task for a landless migrant pauper to become a fully accepted member of a new parish in which s/he could find financial support when needed: 'the right to full membership in a local community, which included poor law benefits, came through work service of at least a year, through

property holdings, and through public service' (Lees 1998: 28). Young apprentices acquired the settlement in another parish if they served an apprenticeship for seven years there (Sokoll 2000: 22). Women, however, settled automatically in their husband's parish when they married (Feldman 2003: 85). According to the members of the privileged classes in charge of the granting and distribution of relief, paupers belonged to one of three categories (Lees 1998: 23):

1. paupers that were worthy and deserving, i.e. industrious people ready to accept any work (even the hardest and basest one), widows, orphaned children, those enjoying a good reputation, and those who were aged, infirm, or struck down by illness.

2. paupers that were in need of work. The parish could give assistance in the job search (Connors 1997: 140) and sometimes subsidised the manual labour with a local farmer.

3. paupers that were unworthy, such as vagabonds, gypsies, drunkards, thieves and rogues. These could be whipped or put away in prison.

Assistance could thus be granted after inspection of the situation of the deserving poor who had applied for relief. It should be pointed out that 'a disproportionate amount of relief was distributed to women [...] since they made up the vast majority of those deemed deserving poor' (Connors 1997: 142). The cases of abandoned wives, single mothers and widows (with or without children) were often investigated, and women were more often than men removed from the host parish (Willen 1988: 562). Even the slightest suspicion of a taste for drinking or of a 'disorderly' or 'dissolute' life could reduce any pauper's chance to obtain relief from the home parish (Sokoll 2000: 27). If the paupers lived in the parish, they could apply for help by going directly to the overseer, an impossible task for those who had migrated to other regions.

Paupers who lived outside their home parish therefore had to send a written request for relief, in spite of their often poor writing skills. Elementary schooling for the poor 'was provided in Sunday schools, charity schools, and dame schools', where they were mainly taught how to read and adopt the right attitude in relation to their inferior social position (Laitinen & Auer 2014: 188). Between 1750 and 1840, 'literacy rates remained relatively stable at *c.* 66 percent of men and 40 percent of women', literacy

meaning the ability 'to mark or sign the marriage register' (Laitinen & Auer 2014: 188). Smitterberg (2012) points out that the term 'literacy' can be interpreted differently, as 'not all people who signed their names were considered able to write other things, and, conversely, some of those who were not able to sign their names were still able to read, since writing was often taught as a separate skill after reading' (Smitterberg 2012: 954). If the pauper did not know how to write a relief request, another person could thus help or even write for her or him (Sokoll 2000). An advocate could also write a request for relief on behalf of the pauper. Relief letters are therefore 'pieces of correspondence written by, or for, individuals and families asking for relief from their parish of settlement when they were unable to make a claim in person' (Shave 2017: 20). Practically, the relief letter was a request for any kind of help from the home parish. The help represented either a weekly allowance, a sum of money or the payment of a debt. It could also be material support such as shoes, clothing, and even fabric for making shirts. Relief requests were generally made when times were hard and the paupers could not find any other solution to their critical situation that was often unemployment, illness or old age. A specific or regular allowance was however not automatically granted, and the paupers had to develop efficient argumentative skills when applying for support.

It is interesting to note that relief letters were written for a specific purpose in an institutional context, i.e. the granting of relief to paupers by landowners, wealthy farmers and other taxpayers from the parish elite and middling sort. The pauper applicants knew their place in the social hierarchy and were also aware of the language constraints linked to their position when negotiating with the authoritative body that could grant them (or not) financial support. Applicants were compelled to address themselves in a particular way to obtain a positive result to their plea. However, each party in the relief process knew her or his place in the exchange and played according to unspoken rules as 'officials, advocates, and poor writers inhabited a space of shared linguistic and referential modes' (S. King 2019: 178). In spite of a relative lack of power over their own fate, paupers had agency and often knew how best to use their desperate situation linguistically to make local authorities act, even though it meant threatening to 'come home' (Sokoll 2000). Rhetoric was thus important when addressing oneself to the authority who had power to grant relief. In his book on the writings of the English poor, S. King (2019) gives a detailed account of how paupers and their advocates used 'particular but

also general rhetorical threads' (S. King 2019: 179) and goes on to argue that '[r]egional dialect and conversational patterns inevitably shaped the expression of language in oral writing' (S. King 2019: 180). But even though the pauper applicant could show great humbleness and politeness, it did not mean that s/he did not know her or his rights, nor that relief would be granted on a regular basis. When making a schematic approach of pauper rhetoric, S. King argues that it mattered who the pauper was, 'whether the letters being considered were written by women, men, the aged, widows, widowers, and so on', because each pauper applicant 'brought his or her own colour and emphasis to common rhetorical vehicles' (S. King 2019: 180). Relief applicants probably knew that if they broke the established rules of how to request relief in their particular situation, they probably risked being denied substantial aid. Some parishes even rejected all claims coming from paupers living outside the home parish (Shave 2017: 127). Efficient rhetoric use, such as expressions of gratitude, submission, respect of the authorities and deference (S. King 2019: 231), as well as the identity construction of 'fellow citizens' and 'fellow humans' (S. King 2019: 185, 189) could prove to be rewarding. For the poor, the Old Poor Law was 'an institutional platform on which the labouring poor could effectively express their needs, pursue their interests and establish their claims' (Sokoll 2000: 46), situation that ended when the New Poor Law was introduced in 1834.

Pauper relief letters have to a certain degree been investigated by historians and historical linguists, but in language history, they seem to be of lesser importance than the writings of the educated elite. So what is special about pauper relief letters? Do these letters vehicle any specific characteristics? And why are they important in language research? In the following subsection, I will discuss the main linguistic characteristics of pauper letters found in the LALP corpus.

2.2.2 Main characteristics of pauper relief letters

As a student assistant in The Language of the Labouring Poor in Late Modern England project (LALP), I have had the opportunity to transcribe many pauper letters and have thus become acquainted with the characteristics of these texts. The literature on standard English during the Late Modern English period (Beal 2004; Auer 2012; Smitterberg 2012; Beal 2016) and writings on ego-documents (Auer et al. 2014; Laitinen & Auer

2014; S. King 2019) as well as on letter writing (Auer, Schreier & Watts 2015; Laitinen 2015) have also given me a solid background to what characteristics might be found in the writings of paupers. My discussion of the characteristics of pauper relief requests is thus based on existing literature but also on my own observations from the transcription process. Apart from the characteristics discussed in this subsection, there are others that would merit a more in-depth discussion, especially the narrative structure and rhetoric used by paupers in their relief applications. For a more thorough discussion on this, the works of Sokoll (2006), Smitterberg (2012), Laitinen & Auer (2014) and S. King (2019) will prove indispensable. In this *mémoire*, I will mainly discuss phonetic spelling, hypercorrection and h-dropping, spelling variation, missing punctuation, the use of upper- and lowercase letters and mechanical writing. I will also briefly mention rhetoric and epistolary conventions.

The pauper petitions from the LALP corpus have certain characteristics in common, features that are interesting elements for linguistic research. As the letters were not written by privileged, educated people using standard 'polite' language but by ordinary working people whose 'writing literacy [...] was gained through a situational mixture of accident, determination, and luck' (S. King 2019: 118), it is the spoken language of the poor that is much reflected in the relief requests and thus represent 'the most authentic records of popular voices' (Sokoll 2006: 91). Even though pauper letters may to a certain degree reflect orality, it is nevertheless clear that the letters do not entirely correspond to spoken discourse. In their letters, paupers might have been led to try to please the recipient through what they believed to be a more standard language, just as children would do to satisfy their schoolmaster (Crowley 2003: 134–135). 'The language of the poor and uneducated was viewed as deviant and defective' (Crowley 2003: 129), which is not surprising, as paupers had access to no or very little schooling and were not often instructed in writing standard language. They were thus not skilled writers for lack of education.

As already mentioned, the spoken language of paupers shine through in the letters, because of the phonetic writing that is very much present in the letters. I will now discuss this feature before continuing with h-dropping and hypercorrection.

Phonetic spelling (oral writing) and spelling variation

Most of the pauper relief requests use phonetic spelling to some extent. Generally, paupers used Anglo-Saxon monosyllabic words, and as they lived in a social reality of oral communication, even foreign loan words were spelt according to known language patterns: 'when they heard unfamiliar Latinate English, they remembered it by Anglo-Saxon word patterns, not by orthography' (Fairman 2000: 72). Paupers might have drawn parallels between their own way of speaking and how they pronounced the words and how they imagined these words could be spelt according to their previous experiences (or the experiences of others) of written texts. This phonetic writing entails that the spelling varies greatly in the pauper letters, and at times, it can even obscure the intentions of the writer. As readers today are more used to standard written English, phonetic writing might seem surprising to them. When transcribing pauper letters, it is thus recommended to read the letters aloud in order to hear the words instead of relying only on the writing. In doing so, it becomes possible to actually hear what the pauper writer tried to convey through her or his linguistic variant, even though the spelling is not standardised.

As already mentioned, members of the lower social classes received only elementary schooling in the 18th century, so writing standard language must have been difficult for them. Fairman points out that '[t]hough paupers must have spoken 'dialect', they wouldn't have thought of writing it. They aimed at Schooled English, but hadn't had enough Schooling to succeed' (Fairman 2000: 75). Certain standard forms of relatively frequent words were thus unavailable in the paupers' repertoire, and so several spelling versions of the same word could appear in the same letter. As an example of this, the personal pronoun 'we' could become 'whe' or 'wee', and 'they' would sometimes be spelt 'the' or 'thay', depending on the language variant of the writer. The word 'children' can also be found with many different spellings (see Table 4.1 in Chapter 4). Depending on the vernacular of the writer, the oral writing can reveal certain linguistic features that the writer probably used when speaking. The letter written by Richard Jones to the overseer of his home parish (HE/EA/1:) clearly reflects spoken language: words such as 'grate' (great), 'farder' (further), 'tal' (until/till), 'gat Battar' (get better), 'wather' (whether), and 'wat' (wet) might reveal much about the pronunciation of the writer. Phonetic spelling thus gives us clues as to how paupers spoke even though there are no recordings of their

speech. Certain corrections found in the letters were made by the paupers themselves and seem to be connected with the oral writing. The writer might have initially written as s/he would have spoken the words but then either overwrote or erased the faulty letter(s) to add more standard spelling, perhaps encouraged by an onlooker such as a friend or a relative.

Phonetic spelling is a frequent characteristic in most pauper letters. But at times, it is possible to encounter a letter in which a more standard language is used. In this case, an important question is whether the pauper wrote the letter herself or himself or had somebody else write it for her or him. The language in the letter might simply reflect the language of the scribe and not necessarily the language of the pauper. This is an element of uncertainty to be taken into account when analysing the pauper letters of the LALP corpus.

Spelling variation, also recurrent in pauper letters, is a difficult issue to deal with in any corpus but especially in a corpus that is intended to be machine readable. As there are several variants of the same word, the machine will not be able to capture all the related words. If we search for the word 'write', it will find only that word with the exact standard spelling and will bypass all the other variants. This is especially problematic for quantitative research. One way of correcting this is to standardise the language (see 2.1.4). From a linguistic point of view, it is, however, interesting to compare the different spellings that are observable in pauper letters, even though it is difficult to know why paupers wrote the same word differently, sometimes even in the same letter. Perhaps the act of writing was so difficult and time-consuming for the applicant that s/he did not realise there was variation in their spelling. Or could it be so that the paupers were conscious of this but that they did not put too much importance on consistent 'standard' spelling because their desperate material situation overshadowed all the rest? Fairman also suggests that '[w]riters seem not to have got the idea that 'refined' English would be more effective than 'vulgar' English' as relief was still granted in spite of writing difficulties (Fairman 2007: 170). Whatever the reason might be, spelling variation as well as phonetic writing would merit more attention because both present interesting opportunities for linguistic research (Vandenbussche & Elspaß 2007), as well as the phenomena of hypercorrection and h-dropping that I will discuss next.

Hypercorrection and h-dropping

As standard English was slowly gaining ground in the 17th and early 18th centuries, 'the discrediting of dialect use was quick and dramatic' (Görlach 1999: 484). Stigmatisation of certain language variants associated with the lower classes became frequent (Elspaß 2011; Smitterberg 2012). It is therefore not surprising that the population from the labouring classes started using linguistic features such as hypercorrection and h-dropping when having to write (Smitterberg 2012: 957). H-dropping is common in vernaculars and according to Smitterberg (2012: 957), it is 'the most important social marker of LModE pronunciation'. Instead of saying 'I have' with the voiceless glottal fricative 'h', the phrase would be pronounced (and probably also written) 'I av'. Even as late as 1912, this linguistic feature is treated in George Bernard Shaw's play *Pygmalion* to illustrate how the vernacular (including h-dropping) of a cockney flower girl distinguished her from the upper-class speech of Received Pronunciation and how this could be 'fixed' through practice of verbal hygiene[12].

Hypercorrection, on the other hand, is quite the contrary to h-dropping. It implies adding an 'h' where none is necessary in standard English. The reasons for hypercorrection can be debated: Romaine (1982: 265) claims that it could be because hypercorrection contributes to prestige, and as '[l]ower middle class speakers typically exceed the upper middle class in their use of prestige norms in more formal styles', also members of the lower classes might be anxious to over-correct themselves. This could explain why there are many instances of hypercorrection in pauper letters. In the letter written by Richard Jones mentioned above, there are several examples of hypercorrection. Jones, the letter writer, explains that 'to of my little Children his Very Hill' (two of my little children are very ill). Probably anxious to write 'proper' standard language, the father of the two sick children adds the letter 'h' both to 'is' and 'ill', thus almost creating a comical effect that nevertheless seems endearing to a modern reader. Later on in the same letter, Jones also adds the same 'h' to 'I am' which becomes 'I ham', and to 'if' that becomes 'hif'. In contrast, there is an instance of h-dropping as Jones writes 'I ope' instead of 'I hope'.

More contemporary research on language differences between women and men often mention the Uniformitarian Principle that states that 'we must assume that whatever

[12] on verbal hygiene recommended for women, see Cameron (1995)

happens today must also have been possible in the past; whatever is impossible today must have been impossible in the past' (Bergs 2014: 80) and the importance of this principle for historical sociolinguistics. In his research on New York speech dating back to the 1960s, William Labov discovered 'that middle-aged, lower middle class speakers tend to adopt the formal speech pattern of the younger, upper middle class speakers' (Labov [1966] 2015: 101). He then continues by arguing that '[h]ypercorrectness is certainly strongest in women' (Labov [1966] 2015: 101) which indicates that women are particularly sensitive to the use of standard forms and therefore prone to self-corrections. These contemporary findings are valid for the 1960s and onward, and as Nevalainen & Raumolin-Brunberg (2017: 131) argue, 'late medieval and early modern Englishwomen did not promote language changes that emanated from the world of learning and professional use, which lay outside their own spheres of "being"'. Otherwise, Labov's Gender Paradox 'was already clearly in evidence' (Nevalainen & Raumolin-Brunberg 2017: 131) in Tudor and Stuart England. Would it then be plausible to presume that this tendency also existed in the late 18th and early 19th centuries among poorly schooled female paupers – could they have used more hypercorrection than pauper males? As it is difficult to be certain of the identity of the letter writer (perhaps the applicant herself or himself or a neighbour or advocate), the question of female paupers being more inclined to hypercorrection might be difficult to prove. What seems to be clear, however, is that both h-dropping and hypercorrection are important features in pauper relief letters and would deserve more attention than I can give them here.

Missing punctuation and upper- and lowercase letters

Apart from the characteristics that I have presented so far, there are also two other important features to be found in the LALP pauper relief letters: the lack of punctuation and the random use of small and capital letters. It is a fact that punctuation is almost totally lacking – there are very few commas and full stops, except perhaps after the introductory 'Sir', at times followed by a / or a). Generally, the phrases in the letters are not separated by any punctuation, which entails that the whole narration continues without any apparent interruptions or pauses. The writer's ideas are thus spilled onto the paper, 'an internal monologue with clauses chained one after the other' (Fairman 2000: 80), seemingly without any coherent structure. In many cases, this leads to a confusion in the ideas:

descriptions and requests are straightforwardly linked with the coordinating conjunction 'and' or nothing at all, which might blur the message. And yet, the message must have been obvious both to the overseer receiving the relief request and to the pauper writing it – it was a request for help from the home parish. Nevertheless, for a modern reader, the lack of punctuation may seem confusing. The message seems to be obscured by the missing punctuation as it can be difficult for the modern reader to make sense of the phrases. For the overseer, however, who might have known the pauper and used a similar dialect, this was probably his daily grind and caused him no further inconvenience.

The second eye-catching feature of pauper letters is the random use of upper- and lowercase letters. Some relief applicants tend to start each new line of the relief request with a capital letter, while others mix upper- and lowercase without any logic at all. At times, certain capitalised letters are drawn quite artistically: these graphs are generally consistent and appear in all the words beginning with that specific letter. In some documents, it is almost possible to draw the conclusion that the pauper wished to give more emphasis on certain words by making them begin with a capital letter, for example 'Dear Gentlemen', as it was important to address oneself politely to the overseers and the vestry members. Other pauper writings might give the reader the impression that all nouns, and even adjectives, are emphasized by the capitalising of the first letter. Surprisingly enough, the pronoun 'I', which is always capitalised today in standard English, is very often written with a small 'i'. All the same, such conventions seem not to have been known by paupers who had had no or limited access to education and standard language.

When transcribing pauper letters, there is often doubt over these graphs: could this be a capital or a small letter? The size of a letter might indicate that it is lowercase, but the form suggests that it is uppercase. Paleography guides are of great importance when transcribers encounter this kind of problem. Sometimes it is however not very clear whether a letter is capitalised or not, and opinions between transcribers might diverge. In such a case, it is essential to take the whole hand-written letter in consideration, and even the other letters written by the same author, if there are any, and to compare the graphs in their different environments in order to distinguish small letters from capital ones.

Pauper letters can be considered as 'the only authentic trace of people who did not form part of our cultural memory via literary texts, pamphlets, treatises, printed speeches and other documents' as they represented 'a "silent minority"' (Vandenbussche

& Elspaß 2007: 146). Mastering the issues of punctuation and the drawing of graphs are skills that develop with the practice of literacy. The poor, lacking education opportunities, were thus denied access to literacy and could thence not always structure their narration in writing. The lack of punctuation, a constant feature in most pauper letters, and the indifferent use of upper- and lowercase letters, might be a consequence of the manner in which paupers accessed reading and some writing skills. Instead of the more elegant cursive hand-writing, paupers were more likely to draw separate graphs to produce sentences, method called 'mechanical writing' by Fairman (2007) (see also Fairman 2011), method that I will present in the next subsection.

Mechanical writing (drawing graphs)

The question of education for the poor and teaching them how to write were quite important political issues for the elite in the 18th century. On the one hand, the members of the upper classes tried to protect their own privileged position in society by maintaining the poor in relative ignorance as well as distinguishing between education for men and that for women (Purvis 1989; Kiełkiewicz-Janowiak 2014; Auer 2015). On the other hand, the Enlightenment period, the French Revolution of 1789 and the rise of a middle class inspired new ideas around education and social class (cf. Vandenbussche & Elspaß 2007: 147). The 19th century saw a tendency towards mass literacy, and social mobility became possible even for the members of the lower social classes (Simonton 2000; Vandenbussche & Elspaß 2007). Nevertheless, the well-meaning and philanthropic middle-class that arose in the 19th century advocated schooling for the poor, not for reasons of equality but in the belief that 'instruction in reading, catechism, manners and a form of work would teach children their place in society, making them more devout, modest and industrious' (Simonton 2000: 185). Reminding the children from the lower classes of their social (inferior) position 'would create better servants and more civilised and duly subservient labourers' (Simonton 2000: 185), which, of course, was an efficient way for the elite to maintain its privileges.

In the late 18th and early 19th centuries, reading and writing were subjects that were taught separately, and as there was gender segregation in schools, even the sons and daughters from lower-class homes followed different curricula (Purvis 1989: 133). When writing was offered to the poorer students, it was often mechanical writing, i.e. learning

how to draw the graphs and combine the letters of the alphabet to form words. This did not include learning how to write standard language, but only how to graphically produce separated letters (Fairman 2011: 40). Pauper relief requests are good examples of this. Generally, each graph is separated by some space, and the use of capital and small letters is sometimes extremely random, as already discussed. Cursive script is rarely seen in pauper relief letters.

Drawing graphs and knowing how to sign one's name were important steps in literacy for everybody, even for members of the labouring classes, as people were expected to be able to sign legal documents such as wedding registers. Even working-class girls and women were taught some writing, probably mechanical writing, at least from the beginning of the Victorian era (Purvis 1989: 141). But mechanical writing is not sufficient in order to write a persuasive relief request – it is also necessary to know how to use rhetoric and epistolary conventions, as I will discuss next.

Rhetoric and epistolary conventions

As discussed above, the poor were not taught to write nor given any schooling that would have allowed them to become fully literate. The members of the labouring classes were probably capable of drawing the graphs of the letters but as already observed, they were not altogether eloquent nor coherent in their narration. Even though 'the primary trend in English epistolary spelling in the seventeenth century is incipient standardisation' and the 'number of variant spelling forms begins to decrease' (Kaislaniemi et al. 2017: 204) from then on, it seems plausible that this concerned particularly those who had received education and knew how to write both letters and literary texts. Even though many guides to letter writing were published in the 18th century, paupers would probably not have had access to such books that encouraged 'clarity, brevity and an immediately appealing conversational tone' (Sokoll 2006: 100). It seems, however, that the poor were not devoid of all general knowledge of letter writing. They knew how to write a request for relief to their home parish (S. King 2019: 183) even though their spelling and narration were not consistent. In many of the LALP pauper relief requests, it is possible to identify a recurring pattern, i.e. the classical model of a letter with an introductory salutation, an appeal to the recipient's goodwill or even excuses for troubling the recipient, a specific request – financial help – and a conclusive greeting such as the phrase 'I remain your

obedient servant' (Sokoll 2006: 100). Rhetoric and epistolary conventions were thus not unknown to the poor. According to Watts (2015), paupers asked for help when having to write relief requests, probably by such people who had already sent requests, for example friends and neighbours, in order to put the odds on their side and to make their cases heard. Even though the labouring poor often found themselves in difficult life situations, they knew their rights, and not being tongue-tied, they could use all sorts of arguments, even the harshest ones, to persuade the overseer to grant some aid:

> Their threatening letters were highly insubordinate. The letters that those who lived away from their parishes of settlement wrote requesting relief mixed strategic threats of costly and unwelcome returns, with finely tuned pleas of hardship and deferential references to their respectability, their self-help initiatives and their attempts to find work. (P. King 2004: 61)

William Fulwood's manual *The Enemies of Idlenesse* dating back to 1568 had given clear categories of addressees, 'namely superiors, equals and inferiors, and in particular ways in which to write to the respective groups' (Auer 2015: 139). In general, such manuals containing good advice on letter writing were produced for the upper classes and middling sort, but Thomas Cooke's manual *The Universal Letter-Writer; or, New Art of Polite Correspondence* published in 1775 also targets the labouring classes by giving a model of a pauper relief request (Auer 2015: 140). Apparently, manuals, grammars and other prescriptive books concerning language use and linguistic recommendations were not available to the poorer part of the population, and thus models were generally not used by them. It is not entirely known where paupers acquired their rhetoric skills and how they became familiar with epistolary conventions, and yet even paupers imitated them by '[reproducing] the classical model' of letter writing (Sokoll 2006: 100).

Now that the main characteristics of pauper letters have been discussed, it leads me to further questions: how do these linguistic particularities found in the pauper letters become important witnesses of language use and language change in the past? And would it be possible to distinguish gender differences in the discourse of female and male paupers based on these characteristics? In the next subsection, I will explain the importance of the LALP pauper relief letters for linguistic research before continuing with a presentation of the LALP corpus.

2.2.3 The importance of pauper relief letters

In England, the 18th century was the period of constant concern for a 'correct' language, and it is during this time that codification and prescription of the language were spearheaded through an important amount of grammar books, pronunciation and vocabulary guides as well as letter writing manuals and dictionaries (Beal 2004; Auer 2012; Beal 2016). Through a standard language (both written and oral) based on the 'polite language of educated gentlemen' (Auer 2012: 940), the English language was meant to attain a perfect form that would not need to be corrected nor fixed, according to influential authors such as Defoe, Swift and Johnson (Beal 2004). Even William Cobbett, himself a farmer's son who was self-taught in grammar and who later became a fierce journalist, was concerned in his political endeavours that 'those who had most to gain from parliamentary reform were unable to state their case because their language was considered "vulgar" and therefore not fit to express an intelligent argument' (Beal 2004: 98). The speech variants and dialects of the lower social classes were considered as 'debased' and 'corrupt' versions of the otherwise perfect and distinguished standard English used by the elite (Watts 2015: 4). Watts thus argues that during the 18th century, there prevailed a '*myth of linguistic homogeneity*', as it was generally believed amongst the educated that 'a language [could] reach perfection and that it [could] be completely homogeneous' (Watts 2014: 595)[13]. Nevertheless, this ideal of linguistic homogeneity proved to be hard to achieve, as opinions on how to maintain a stable standard language varied greatly (Beal 2004: 91–92), especially as the major part of the English population (about 70%) received no or very little formal education, as already mentioned.

The language in pauper relief requests does not correspond to this ideal of a unified and standard English language. As already mentioned, they present much phonetic writing and spelling variation as well as a lack of punctuation. It is a fact that women and members of the lower social classes wrote less than the educated male elite (Smitterberg 2012: 953–954) and their writings have not found a place in language history. Therefore, pauper relief letters are interesting research objects as much linguistic exploration has so far been led on the development of standard English, and not so much on non-standard writings. Creating (electronic) corpora of archive material is a costly un-

[13] see also Watts 2015

dertaking and a time-consuming task, which might explain why certain documents have not yet been researched. The LALP corpus, the collection *Essex Pauper Letters* and other documents that represent the language of the labouring classes are therefore important materials 'to catch a glimpse of the lives, fates and language use of ordinary people' (Laitinen 2015: 201) in the late 18th and early 19th centuries. Pauper letters do not, however, give scholars access to the spoken language of the past. Nevertheless, they are speech-like texts because they present much phonetic writing, and they can serve as reliable sources for understanding Late Modern English pronunciation in the late 18th and early 19th centuries (Smitterberg 2012: 953). Researching this material originating from the artisans and labouring poor might permit a deeper insight into and a more varied picture of the language use and language change related to the English language. The fact that paupers sometimes were unable to write themselves and turned to family and friends for help instead of hiring a professional scribe does not disqualify these documents as linguistic sources. As Sokoll argues,

> [letter writers] often go far beyond mere strategic considerations, and open deep insights into the everyday life of the labouring poor. Moreover, in linguistic terms, pauper letters 'sit' closer to the experiences and attitudes of the labouring poor than most other records. Some of them, especially those written with heavy phonetic spelling, may almost be regarded as 'oral' testimonies. The fact that pauper letters were not necessarily all self-written does not invalidate this. The important point here is rather that, among the hundreds of different hands in evidence in the record, there are hardly any hands of professional scribes. (Sokoll 2000: 29)

Furthermore, Sokoll highlights the fact that '[p]auper letters are of major importance for the social history of poverty from below, since they provide – literally – first-hand evidence of the experiences and attitudes of the poor themselves' (Sokoll 2000: 25). The LALP pauper letters are thus interesting witnesses of spoken discourse because they contain a considerable amount of phonetic spelling and non-standard language and reflect the everyday-life conditions of the vast majority of the English population during the Late Modern English period. Researching pauper letters and documents written by members of the lower social classes is thus 'of great value because socio-historical linguists now have the possibility to compare language use in the Late Modern English period across all social levels' and it will 'allow for more accurate descriptions of linguistic vari-

ability in Late Modern England' (Auer et al. 2014: 25)[14]. So far, language history has been established from 'above', and through the written documents of members from the lower social classes, it is now also possible to investigate language from 'below', not representing standard language (Elspaß 2011).

2.2.4 Description of the LALP corpus

Today, there is much interest in research based on documents representing 'language from below', for example through ego-documents such as pauper letters and personal diaries of emigrants (Hitchcock 2004; Elspaß 2011). Turning to a wider variety of text types than earlier for linguistic research seems interesting not only because it is now possible to use enormous quantities of texts through electronic corpora and corpus linguistics, but because 'the written record available is also strongly biased toward formal writings of highly educated men from the upper ranks of society' (Auer et al. 2015: 5). Over these two or three past decades, it has become evident that so far, the historical documents used for linguistic research in English largely represent the language of the English elite that 'consisted of less than 100,000 people' out of a population of about ten million citizens (Auer et al. 2014: 10). This might give rise to bias in the research conclusions as the examined material does not necessarily take into consideration the language of the labouring poor: '[i]n this "language history from above" approach, the histories of non-standardized languages and language varieties were widely ignored' (Elspaß 2011: 3). Apart from the neglected writings emanating from the lower social classes, or non-gentry classes, historical writings by women have also been overlooked for different reasons (Smitterberg 2012: 955). If such documents that have been of limited interest to scholars so far would be included in linguistic research, it would most probably give rise to new research questions as well as an enlarged vision of the evolution of the English language (Auer et al. 2014). In this context, creating electronic corpora of historical documents from the labouring classes, such as the LALP corpus, and using them for linguistic investigations would be of great importance to obtain a more nuanced view on language use and language change.

[14] see also Auer, Schreier & Watts 2015

The Language of the Labouring Poor in Late Modern England project (LALP) is a four-year Swiss National Science Foundation research project that started in 2020. Based on the independent researcher Tony Fairman's important collection of about 2'000 pauper petitions written approximately between 1795 and 1834, this project aims at 'gaining a better understanding of the role of social stratification in real-time linguistic change and at complementing the "traditional" history of written English during its late codification and early prescription stages, particularly the period c. 1780-1840' (https://wp.unil.ch/lalp/project-description/). Fairman's initial intention was 'to investigate lower class writing' as well as 'to capture as many of the physical properties of handwriting and the social backgrounds of the applicants as possible' (Auer et al. 2014: 12). The LALP project continues on this line and goes even further. Its aim is to create a 'corpus that could serve an interdisciplinary research community' (Auer et al. 2014: 12), for example historians, sociolinguists, historical sociolinguists and computational linguists. Today, the LALP corpus represents more than 2,050 letters written by paupers applying for out-parish relief from their home parish.

As Fairman worked on the transcriptions for about 20 years, 'his transcriptions and the coding used were not necessarily suitable to be transferred to plain text versions' (Auer et al. 2014: 12). This is why extensive work on the transcriptions is currently being done by the LALP team members who are developing this important collection of pauper relief letters into a formal corpus. All the different steps and processes of transforming these hand-written historical documents emanating from the lower social classes into an electronic searchable corpus are also being undertaken. These processes include work on the transcriptions (comparing earlier transcriptions with new transcriptions, elaborating missing transcriptions and cross-checking transcriptions), generating plain-text versions and adding metadata, examining authenticity, proceeding to spelling normalisation and converting the material into a Historical TEI-compliant format. The electronic corpus will give not only linguists but also historians and researchers from related disciplines the opportunity to explore these manuscripts. The project will be completed in 2024 in spite of certain practical difficulties encountered during the 2020 pandemic. Once fully operational, the LALP corpus will give scholars and researchers new insights into the language use of the members of the labouring classes as well as into the linguistic strategies used by paupers when applying for out-parish relief.

After having discussed the corpus as a tool in linguistic research, pauper letters and their characteristics as well as the LALP corpus project, I would now like to move on to the question of gendered discourse in relief letters. Many interesting questions arise from what has been so far discussed: if the letters were thoroughly analysed, would it be possible to detect any linguistic differences between the language in letters written by female and male paupers? If so, what differences could be uncovered? And what reasons could be found for such differences? In the next subsection, I will introduce these questions before continuing with a qualitative and quantitative analysis of a limited subcorpus of ten letters written by female paupers and ten by male paupers.

2.3 Gendered discourse in relief letters?

From the 1960s onward, the differences between the speech of women and men have been looked into, with mitigated success and controversial results. According to the Uniformitarian Principle that claims that 'the processes which we observe in the present can help us to gain knowledge about processes in the past' (Bergs 2014: 80), it could be argued that gendered discourse distinguished today might have roots in the past. Gendered discourse may thus stem from earlier centuries and could have tainted language already hundreds of years ago. Would it then be possible to find traces of these divergences in the LALP pauper relief requests? Did paupers use language differently depending on their gender?

Before discussing previous linguistic research around the language of women and men and the exploration of gendered discourse in the LALP pauper petitions, it is necessary to make a historical journey of women's lives, social conditions and access to education in earlier centuries. Being aware of certain sociohistorical facts is of great importance in language research, and perhaps even more importantly so when it comes to gendered discourse. As we have seen, pauper letters may show phonetic writing, inconsistent spelling and several other characteristics that were long considered as typical for the language of the labouring classes. The lack of access to the written standard language was thus a reality in England for most members of the lower classes up until the Education Act of 1870 (Fairman 2011). It was generally the male population from the higher social classes who had access to classical education while girls were generally excluded from the learning processes reserved for boys (Nevalainen & Raumolin-Brunberg 2017: 40). It

may thus be argued that 'the upper layers of society and men are over-represented in diachronic corpora' (Auer et al. 2014: 10) and that this needs to be taken into account when researching historical documents and texts. This means that 200 years ago, education was socially stratified and there were literacy differences between women and men of all social classes. The social context was also considerably different as women did not have the same civil or political rights as men (Nevalainen 2000: 39). As already mentioned, formal education was the privilege of such families that could afford to send their children, or rather sons, to private schools. It is important to underline that historically, boys were systematically favoured in educational questions compared to girls, no matter the social class they belonged to[15]. Patriarchal ideologies that 'identified [women] with the home and family and men with paid work and economic power' (Purvis 1989: 53) developed early on but became particularly present in the Victorian era. In all times, women from the upper classes were not supposed to work and were therefore rarely given any serious education, except such education that could be useful in the private sphere of the home (Purvis 1989). Women from the working classes were expected to work, either as servants, maids, field hands and seasonal labourers, or as assistants to their husbands in their professional activities such as shops and workshops or in trade. Even for working-class women, 'social status came from a woman's family position in the social hierarchy, not from her own skills as a professional' (Nevalainen 1999: 511). Women were thus totally dependent on a male relative – either a father, brother or husband – which meant that without any professional proficiency nor education, marriage and continuous dependency were thus the only solutions for women from any social class (Fletcher 1995).

Patriarchal ideologies and societal structure thus kept women in an inferior social position and confined to their homes. In the Late Modern period, women from the middling sort – as wives and mothers – only received such education that was 'ornamental', and yet they were responsible for providing an ideal of moral power to others (Nevalainen 2002: 186). Later on, in the Victorian era, even women from the lowest classes were relegated as much as possible to the private sphere of the home because of their biological capacity to bear children which made them 'generally responsible for the social condition of society' (Purvis 1989: 55). The woman was seen as a man's property

[15] see Purvis (1989) for a detailed description of working-class women's participation in educational institutions in the 19th century

that could be transferred from her father to her husband (Purvis 1989: 50), and the education of working-class women of the 1800s 'was justified in terms of her future life as a practical, efficient housewife who served the interests of a husband, children and society' (Purvis 1989: 142). In the 18th century, standardisation of the English language as well as linguistic prescription were well on their way, which means that women from all social layers must have known that if their children (especially their sons) were to have a better and more prosperous life by climbing the social ladder, they needed to speak and write 'correct' English to avoid being ostracised (Smitterberg 2012: 953, 961). And yet, the voices of women were rarely heard. In general, women in early modern Europe were prescribed silence as verbal hygiene, at least from the 17th century onward (Cameron 1995: 172–174). Women were already then publicly encouraged to speak little, to be solitary and show verbal obeisance when speaking to their men. Gradually, 'this silent and domesticated woman became the dominant ideal of linguistic femininity for society as a whole' (Cameron 1995: 174).

And yet, there is proof that the poor made their voices heard through writing. Both female and male paupers knew how to write and compose for example basic letters, even though they did not have access to standard English through dictionaries, guides and manuals in the same way as the elite. In the 18th century, schools attended by the poor put the main focus 'on religious discipline and social subordination' for both working-class boys and girls, while girls learnt to read, write and sew while boys had access to reading, writing and accounts (Auer 2015: 138). In the 19th century, women were only reluctantly admitted to mechanics' institutes and 'male and female students and women of different social classes were offered different curricula' (Purvis 1989: 128). The curriculum of women could thus be biased according to patriarchal ideologies and could include 'messages about women's traditional place within the sphere of the home, the political ideas of the nineteenth-century women's movement, and images of women's "public" role' (Purvis 1989: 128). Gender segregation was common in mechanics' institutes, with separate classrooms for women and men, and at Pudsey, there were even 'separate nights for attendance', with only two evenings a week for women (Purvis 1989: 129). Class segregation also existed: 'curriculum was differentiated by social class' (Purvis 1989: 133) as the cheaper evening classes were intended for working-class women and the day-classes,

more expensive, for the middle-class women. The differences in curricula between female and male students is described like so:

> Gender divisions between women and men were already well marked by the late 1850s and the 1860s. At Lockwood in 1857, women were taught the three Rs, grammar, knitting, sewing and marking, while the men could study the three Rs, algebra, mensuration, history, geography, grammar, music and freehand and ornamental drawing. (Purvis 1989: 145)

Working-class girls and women were thus clearly disadvantaged as far as schooling opportunities were concerned. These facts lead me to certain questions concerning possible gendered discourse in pauper letters. Since men from all social classes, also the working-classes, were generally better schooled, would it be plausible to presume that male paupers were advantaged compared to female paupers when it comes to writing relief requests? Or did all paupers, males and females, have identical (dis)abilities in letter writing, as this was not necessarily taught at school? Perhaps male paupers had better access to standard written English than female paupers when addressing themselves to the overseer? And if this is so, does it entail that there are any significant gender differences in the language of the LALP pauper letters, for example in style, vocabulary or grammar? The difference between the language of the educated women, who had had access to the standard spoken and written language of the elite, and that of the working-class women, who resorted to sporadic life-long learning, can be proved to be considerable. As Auer (2015) argues, 'stylistic variation in Late Modern English letters is largely determined by the written linguistic repertoire one was able to accumulate by way of formal and informal instruction as well as by self-improvement and practice' (Auer 2015: 154). But between female and male paupers, it is, however, difficult to know exactly how much their discourse differentiated and whether women and men used language differently. As Kiełkiewicz-Janowiak (2014) points out, it is a challenge to reconstruct a relationship between gender and language use with historical data, especially when data are missing, for example the degree of education that each pauper had had access to.

During the latter part of the 20th century, sociolinguistic patterns and the language of women and men have been studied by for example Fishman (1968), Labov (1972) and Lakoff (1973) and many more. It is nevertheless questionable whether it is

possible to directly apply modern language theories of gendered discourse on historical testimonies of language. To deal with gendered discourse in the past, it is necessary to turn to historical sociolinguistics that considers 'the interrelatedness of linguistic and social factors' (Auer et al. 2015: 2) in order to avoid any possible pitfalls concerning language use in past centuries. Nonetheless, it is interesting to compare research made on gendered discourse in contemporary English and the findings of research applied to earlier periods, as modern theories might give clues to former language patterns. According to research on contemporary speech, women are generally considered as more willing to use linguistic forms that carry prestige and that correspond to a more standardised language (Labov 1972; Romaine 1982; Labov 2001) even though 'men also seem to be responsible for the introduction of new norms, although these are usually vernacular norms which are said to have "covert prestige"' (Romaine 1982: 264). Labov (2001: 267) calls this apparently female inclination a 'conservative tendency of women' and goes even as far as arguing that it is a proof of 'women's superiority to men in all aspects of verbal behavior' (Labov 2001: 291). According to Labov,

> both conservative and innovative behaviors reflect women's superior sensitivity to the social evaluation of language. In stable situations, women perceive and react to prestige or stigma more strongly than men do, and when change begins, women are quicker and more forceful in employing the new social symbolism, whatever it might be. (Labov 2001: 291)

Labov (2001: 292–293) thus argues that women, especially those from the lower middle-class, are leaders of language change as he discusses this Gender Paradox (or also called the Conformity Paradox). In her article on sociolinguistic variation in Early Modern English, Nevalainen also discusses the fact that '[s]ociolinguistic research shows that linguistic innovations are typically transmitted in informal conversations, by young people, middle-class speakers, and women' (Nevalainen 1999: 499). In consequence, social class and gender seem to have their importance when it comes to the adoption of prestige forms in spoken language (see also Romaine 1982: 265). To corroborate this, further research on gendered discourse seems to indicate that already in the Early Modern English period, women were those who drove language change since they were apparently more prone to adopt high-frequency variants, thus gaining 'a supralocal status' compared to men who had a tendency to adhere to 'strictly localized linguistic forms' (Nevalainen

2000: 38). As already mentioned, Nevalainen & Raumolin-Brunberg (2017) argue that Labov's Gender Paradox is correct for the language of women in the Tudor and Stuart period, except for such linguistic spheres in which they were not present.

As we have seen, recent linguistic changes in spoken English are spurred on by women and members of lower social classes in our contemporary society (Labov 2001). If we consider the Uniformitarian Principle, could it then be argued that women adopting linguistic innovations is a historically repetitive pattern, and that what Labov and Romaine suggest is equally true for women from the working-class in earlier centuries? Could this mean that 18th-century female paupers that were only partly schooled would have been more inclined to use standard language forms advocated by a social elite in their oral communication with the overseer as well as in their relief requests, as they would have been particularly sensitive to verbal prestige and stigma? And how can we be certain that such a tendency can be assigned to women, as the identity of the letter writer is often unknown? Since the English language was in the process of being codified in the 18th century by the educated elite and the poorer part of the population received only limited schooling, it does not seem plausible that 18th-century working-class women were the initiators of language change. Interpreting the past through the present undoubtedly requires cautiousness, as for example social factors need to be taken into consideration, especially when it concerns a population that had no or little opportunity to receive schooling and formal education.

All researchers do not agree that gendered discourse truly exists. Biber, Conrad & Reppen (1998: 216) argue that 'there is surprisingly little empirical research' on the issue of women and men talking differently. As they continue the discussion from a historical point of view, they also question whether there have been 'systematic differences in the language of women and men during different historical periods' (Biber, Conrad & Reppen 1998: 216). If such systematic differences existed, it would be pertinent to question whether 'the relationships between the language of men and women remained constant across these periods' (Biber, Conrad & Reppen 1998: 216). It could then almost be argued that if there are instances of gendered discourse today, it might probably be due to historical reasons and that gendered discourse might have been present in earlier days as well. To support this hypothesis, we can apply the Uniformitarian Principle of sociolinguistics, stating that 'the fundamental principles and mechanisms of language

variation and change are valid across time' (Auer et al. 2015: 4). It is nevertheless not possible to know exactly to what extent there was gendered discourse disparity as a vast majority of the historical texts that have survived and are known to us emanate from a male power elite who wrote standard English. Relief letters written by paupers such as the ones present in the LALP corpus are therefore valuable for researching gendered discourse and making comparisons with already available data.

As already seen, the LALP pauper relief letters reflect certain issues that are directly linked to missing information about the writers and the so-called 'bad data' problem (see 2.1.6). Gathering historical facts around language can sometimes prove to be an almost impossible task as we simply do not have access to historical background information or historical evidence (Labov 1994; Auer et al. 2015). As already mentioned, historical texts such as records of trial proceedings and witness depositions might seem to mirror spoken language, but authenticity cannot always be proved since language transferred from spoken to written form may be amended by scribes and judges 'to support a particular religious or political bias' (Kytö & Walker 2003: 241). Therefore it is important to be aware of the fact that 'there is always the degree of scribal, as well as editorial, interference to consider' (Kytö & Walker 2003: 241). Exploring gendered discourse in the LALP pauper petitions might nevertheless reveal certain answers to questions of language use of female and male paupers. In such explorations, it is essential to approach these data with care, also taking socio-historical aspects into account:

> To further the study of historical sociolinguistic variation it is important to add socio-historical depth to flat demographic dimensions, increase awareness of the social aspects of these dimensions, and make students of language understand that it is indeed these aspects that have consequences for language use patterns. One way to tackle this task is to read the historical text to understand it (and others like it) from its own perspective and through its own discourse. (Kiełkiewicz-Janowiak 2014: 307)

In modern-day England, such concepts as social class, Marxist theory, gender inequalities, computational models and Equal Pay Day are well-known. Such modern concepts and the discussions around them did not exist at the end of the 18th century, and there is therefore a risk of distorting the past with our contemporary perspective. As we

'transpose modern concepts such as *social class*, *gender* or *prestige* to historical settings', we are in danger of creating important 'pitfalls of anachronisms' (Auer et al. 2015: 5). It is known to us that life was different in the 18th and 19th centuries but we do not know exactly how different it actually was (Labov 1994: 11). Nevertheless, the issues of anachronism can be avoided through interdisciplinary work:

> it is the task of historical sociolinguists to reconstruct a broad picture of the social context in which the language varieties under investigation were used, drawing on the inductive method to identify the social conditions of language variation and change, ensuring empirical, social and historical validity. (Auer et al. 2015: 5)

The inductive method based on sociohistorical facts and interdisciplinary approach can undeniably help researchers to establish new hypotheses around language change even though there is a problem of 'bad data' as well as uncertainty relative to the historical documents and related facts.

Apart from the necessity to anchor texts in their time period and social context, there is another fundamental requirement for a thorough exploration of gendered discourse in the LALP pauper petitions, and that is a completely functional electronic corpus of the hand-written pauper relief letters. At the moment of writing, the transformation to a machine-readable corpus is still in progress, and therefore it will not be possible for me to make an exhaustive analysis of all the LALP pauper letters, which, by the way, would go far beyond the scope of this *mémoire*. My intention therefore is to create a subcorpus in order to compare ten letters written by female paupers with ten letter written by male paupers. The language of the transcriptions will be normalised with the VARD2 program and the results will be analysed qualitatively as well as quantitatively with the help of AntConc. A description of these different steps will be provided in sections 3 and 4.

Chapter 3

Methodology

So far I have led a reflection on the importance of the corpus as a model of historical linguistic phenomena, in this case gendered discourse, as well as possible pitfalls and problems of uncertainty, bias and the so-called 'bad data' problem linked to historical documents and historical data in general. I have also presented the LALP corpus that is currently being designed and the pauper letters that constitute the corpus. At this stage, I would like to treat the question of possible gendered discourse in the pauper letters.

As I have already mentioned, a complete analysis of possible gendered discourse in the pauper letters of the LALP corpus would represent an overly ambitious task in this *mémoire*. Nevertheless, my aim is to try to find indications of gender differences in the language of pauper letters, or simply find clues to such differences. By carrying out a qualitative and a quantitative analyses of a subcorpus consisting of ten relief letters written by female paupers and ten letters by male paupers, I aim to shed new light on gendered discourse. It is nevertheless important to keep in mind that these explorations might not lead to any reliable conclusions about gendered discourse in the LALP pauper relief requests. As my subcorpus is limited to twenty letters representing ten female paupers and ten male paupers, the results obtained might not be valid for the entire LALP corpus, or only partially. The analyses might nevertheless give a general idea of what could be found when investigating gendered discourse in lower-class relief requests sent by paupers to their home parishes.

The different steps in my exploration of the subcorpus are the following: (1) the choice of ten pauper letters written by women and ten others written by men accord-

ing to certain criteria determined beforehand; (2) a qualitative analysis of the plain-text versions focusing on certain linguistic features; (3) the spelling normalisation of the letters with VARD2; and (4) a quantitative analysis of the standardised letters with AntConc. Basing my research on these steps, I will try to formulate a hypothesis on possible gendered discourse in the LALP pauper letters.

The twenty letters of the subcorpus were chosen randomly among the LALP pauper letters with existing metadata. It was difficult to find letters where authenticity was certain, i.e. written by the pauper applicant. I therefore decided to include such letters that are considered as authentic and (likely) autographical. In my choice of letters, I respected two criteria: gender of the sender and petitioner role as mentioned in the metadata of each letter, and randomness of county. I chose such letters where the metadata indicate that the gender of the sender and of the applicant is the same, either <SG F> and <AG F> for women or <SG M> and <AG M> for men. There is one exception to this – the letter sent and signed by Stephen Orrill (NG/MA/1) that indicates <SG M> and <AG X>. One letter (HU/BR//3: 8+) is likely non-autographical ((N) 2) as several letters from the same sender exist but are written in different hands. The metadata of two letters gave no indication of the authenticity status of the letter (DU/BC/3, DU/BC/5). In spite of these facts, my decision was to include these letters into my subcorpus. As can be seen, the difficulty with pauper relief requests is the authentication of the writer. Did the applicant herself or himself write the letter or did someone else write the letter for her or him? Did the letter writer apply for relief for herself or himself, or perhaps for a child or for the whole family? Perhaps the recipient of relief was a friend, and the writer applied in her or his name? And did the pauper herself or himself sign the letter, or was it the writer who did that? Was the letter sent by the applicant herself or himself or by someone else? There are many questions to which there are no answers. But as S. King (2019) argues, paupers had to write efficiently as 'the groundwork for the negotiative process had to be laid subtly' (S. King 2019: 89). It seems that in spite of limited writing skills, paupers succeeded well in this as they were granted relief, whether they wrote themselves or were helped by others.

As for the second criterion, the letters were chosen randomly from different counties as I did not wish to limit my choice to only one county or to impose the choice of one letter per county. For certain counties, there is a large choice of letters already

provided with metadata. As the work on the corpus is ongoing, metadata is lacking for a certain part of the corpus. Therefore, there are up to three letters stemming from the same county for the letters written by female paupers and two for those written by male paupers.

In the choice of letters, I did not consider the length of the relief letters. Some letters of the subcorpus are very short, presenting less than 100 words, while others are longer, containing up to 400 words or more. Most pauper relief requests seem to be quite short, at least those that I have so far encountered in the LALP corpus. The letters generally contain about 200 words or less and are written on one page, continuing sometimes on the back of the same sheet. In my subcorpus, there are therefore shorter letters as well as longer ones, some concerning the applicant herself or himself, others concerning the family (parents or children of the writer).

When creating the subcorpus, I did not take into consideration the number of letters written by the same applicant to the home parish. Some letters might be part of a longer series of exchanges, ranging over several years, while others are individual letters, the applicant appearing in the LALP corpus only once. As far as the hand-writing is concerned, some letters are written by a less experienced hand while others mirror a more experienced letter writer capable of a more efficient narration and epistolary structure.

The plain-text versions available in the LALP corpus are linked with metadata about the pauper letters and contain the original spelling of the letters together with transcription annotations indicating for example holes, tears, folds, damages, insertions, and words that have been rubbed out or crossed out. I eliminated all these annotations and kept only the actual text of the pauper letters. Certain words that were illegible were left out, and such words that were uncertain but fairly obvious were added in full.

For the qualitative analysis, the focus was put on five aspects of the plain-text versions: the address form at the beginning of the letters and the final salutations, lexical differences (or similarities), punctuation, traces of standard writing, and finally the oral features, or phonetic writing. Politeness would have been an interesting feature to explore but this topic proved to be too ambitious to be treated in this *mémoire*. Even though I will not discuss this topic in more depth, I will, however, briefly mention the question of politeness in relief letters in connection with other linguistic elements.

In order to be able to carry out a quantitative analysis of the pauper letters with AntConc, it was necessary to transform the cleaned plain text versions into texts with normalised spelling (modern standard English). As mentioned earlier, the relief letters present a great variety of spelling, which makes it challenging to use them in their original form with computational techniques such as CL. As a consequence, it is advisable to transform the phonetic spelling into standard English. To normalise the spelling, I used the computer program VARD2 (see 2.1.4) which was developed mainly for Early Modern English texts and has already been used for normalising the language of documents from that period. VARD2 can also be used to normalise texts from other historical periods, such as the Late Modern English period.

The linguistic variables that are discussed here, both for the qualitative and the quantitative analyses, were agreed upon with my supervisors and are based on my personal experiences of pauper letter transcriptions as well as on previous literature on the Late Modern English period (Fairman 2000; Fairman 2007; Smitterberg 2012; Auer & Fairman 2013). As already mentioned, the quantitative analysis was performed with AntConc. In this analysis, I used the standardised texts of the pauper letters. First, I did a word count and calculated the mean for letters written by female paupers and male paupers (see Table 4.2). Then, I made word lists for both genders and compared these lists. The results of the word lists encouraged me to look more closely at the vocabulary of female and male paupers and at certain specific words, such as the personal pronoun 'I', 'that' and the presence of Anglo-Saxon versus Latinate words. The important difference in the use of *that*-clauses between women and men encouraged me to explore this issue further.

Chapter 4

Presentation of results

4.1 Qualitative analysis

As a general remark, I would like to point out that while exploring the plain text versions of the LALP corpus, I noticed that a great many letters were written by women. This could of course be a pure coincidence. Surprisingly enough, in some counties, most of the LALP pauper letters were actually coming from female paupers who were either widows (often with children) or women writing for their family, the husband being ill or otherwise prevented from addressing himself to the overseer. This is an ambiguous finding, since in the late 18th and early 19th centuries, it was generally the task of the head of the family (husband or father) to be in charge of the affairs with the officials and local authorities such as the home parish, as already mentioned in the section on women's position in the patriarchal society of the 1700s and 1800s (see 2.3).

S. King (2019) draws our attention to the fact that in the corpus of letters he researched, '[t]he majority of all letters, some 68 per cent, were sent by or for men' and goes on to specify that 'this concentration probably reflects the fact that they applied on behalf of other household members rather than arising out of differential literacy rates or need' (S. King 2019: 28). I assume that S. King bases his findings on the signatures contained in the letters, as well as in-text references to men applying for relief. The fact is that married women could write relief requests and sign with their husband's name (S. King 2019: 33–34), which makes it very hard to identify them as the letter writer. If the signature only is trusted, then the conclusion might be that a majority of men wrote relief requests. Identi-

fying the wife as the writer then becomes difficult: it is only through either in-text references and/or regular correspondence showing the same hand but with the wife's own signature that it becomes possible to find the true identity of the writer. Furthermore, the fact that men often wrote relief requests does not necessarily prove that women did not know how to write relief applications to the overseers. It seems as if women would apply for relief whenever their husbands were incapacitated to work because of illness or injuries, or because the woman had been abandoned or widowed and had no close male relative to step in for her.

In the LALP corpus, there are also applications made by widows for their children, and here it is plausible that they would be the letter writers. Being widows, they were considered as deserving paupers, which could be one reason to put forward the marital situation as an aggravating fact leading to poverty and distress. The exposed socio-economic situation of a widow apparently justified the need for relief. S. King (2019) argues that '[i]t is now well established in the British and European context that widows were disproportionately represented in the highest and lowest socio-economic groups of both urban and rural communities' (S. King 2019: 296), which might explain my first impression of women being frequent relief applicants. Therefore it would be interesting to explore the entire LALP corpus to investigate whether the number of letters coming from widows is in disproportion to the rest of the letters. As a matter of fact, there are four (perhaps five) widows applying for relief in my subcorpus (CA/TR/2:31, DU/BC/3, DB/BK/1: 4, DB/TI/2: 12r, OX/CL1: 1) as well as two women who seem to live on their own because of old age (BF/SH/2) or for some other reason that is not mentioned in the letter (HE/BR/1: 9+).

An example that slightly contradicts King's argument about the majority of letter writers being men is that of the local authorities of Hackney that dealt with 99 cases of relief requests between October 1731 and August 1753. In this case, the majority of requests were either written by women or concerned also women in the family:

> 18 appeals were made by males; 37 were made by single females; 30 were made by women with children; 12 were made by families; and 2 were made on behalf of children. Not only were single women (this includes widows and spinsters) the highest category, but by combining all the categories that included women, 80 per cent of the cases involved women directly. In the majority, 67 per cent, women

> were the principal appellants. These statistics are remarkably similar to calculations made by other scholars seeking to gauge the scale of pauperism amongst women in the early modern period. (Connors 1997: 139–140)

As can be seen, the results may depend on the formulation of the research question: Are we looking for pauperism amongst women and interpreting the relief requests according to certain criteria that would not be the same as if the main goal was to identify letter writers on the basis of the signatures present in the letters? Both these approaches are valuable and certainly worth pondering and researching. It is therefore important to carefully interpret the results of any analysis based on data that present uncertainty and includes different socio-economical factors that we are not wholly acquainted with because of the time lapse between now and then as well as the 'bad data' problem.

After this general remark, I would now like to continue with the qualitative analysis of the LALP subcorpus by discussing the following points: the address form at the beginning of the letters and the final salutations, lexical differences (or similarities) between female and male relief applicants, punctuation, possible traces of standard writing, and finally phonetic writing. The numbers in brackets refer to the LALP pauper letters.

The address form at the beginning of the letters and the final salutations

Both female and male paupers start their letters in equal proportions by addressing the recipient or recipients with either 'Sir' or 'Gentlemen'. In the subcorpus, there is no comma after these words. After 'Sir', there is either nothing, a slash, /, or a bracket,). Two female paupers use this (CA/SW/5: 10, HE/BR/1: 9+) as well as two male paupers (NG/MA/1, OX/CH/1: 4). If the applicants knew the overseer personally, they could also mention his name, for example 'Mr Martin Sir' (HU/BR/3:8+). For relief requests coming from paupers, it does not seem common practice to start with 'Dear Sir' or 'Dear Mr Martin'. The reasons to this might be numerous: The paupers did not necessarily know the overseer personally, or perhaps it simply was not customary to use such a formulation with the overseer but more in use with family members, for example 'my dear Wife'.

'Gentlemen' also seems to have been used when the applicant did not know who s/he was writing to. Paupers knew that overseers consulted other members of the

local community in order to take their decisions about relief, and addressing oneself to several people might have been a sign of modesty and respect. A letter could also start with 'to you Gentlemen and overseer of...' (DB/BS/6: 1) or directly with 'to the overseers of...' (DB/BK/1: 4). In my subcorpus, one male applicant (DB/BK/4: 15+) goes as far as to write 'Sir I make bold in adreſsing a few lines to you' when he writes on behalf of his parents, formulation that gives the impression of a humble yet brave man, facing his superiors in such a delicate request.

Immediately after the first salutation, the applicant often asks for forgiveness for writing or asks the recipient not to be offended by the letter. Standard phrases such as 'I have to inform you', 'I am sorry to be troublesome but...', 'I am sorry to be forced to trouble you with these lines' and 'I am compelled to write to you' are much used, and both female and male paupers adopt these. Women tend to show a very humble attitude by either thanking for the previous payment (CA/TR/2: 31), by stating 'it is with great pain that I have to inform you...' (DU/BC/3), 'I hope you will excuse me taking the liberty of writing' (DU/BC/5) and 'I am sorry to inform you that I am obliged once more to trouble you' (BF/SH/2). Men, on the other hand, seem more prone to immediately explain facts, such as a wife's or a child's illness (DB/BS/6: 1), a meeting with either the person who transmits the relief money or the host parish overseer (NG/MA/1), or the description of debts (HU/BR/3: 8+).

As for the final salutations, there is a striking difference between the female and the male pauper petitions. Most female paupers in my subcorpus embrace an attitude of what I would like to call polite submission, and apart from one exception (HE/BR/1: 9+), the female applicants end their letters with the standard polite phrase 'I am your humble servant'. This phrase was part of the epistolary frame that even aristocrats used (Fitzmaurice 2015: 168), and it seems surprising that female paupers should use this more than male paupers. Some women even accentuate their respect by saying they are 'humble and obedient' (DB/TI/2: 12r) and 'humble and dependent' (DU/BC/3). Others ask for advice what to do in their situation (OX/CL/1: 1), express hopes of receiving something without delay (CA/TR/2: 31) and beg for a favour (DU/BC/5). One female applicant even goes as far as saying 'yours to command' (DB/BK/1: 4). The role, age and social position of the addressee seems nevertheless of importance. In aristocrat letters, 'the selection of opening compliment, ritual expressions of gratitude and expressions of friendship ap-

pear to be conditioned less by the motive in writing and more by the relative seniority of the addressee to the writer' (Fitzmaurice 2015: 168). For paupers, both motive and social position of addressee seem to have been of importance. Paupers used the same epistolary frame but wrote their letters in quite a different situation than aristocrats, as their main goal was to receive the advantage of financial help from their parish of settlement.

Male paupers, however, do not show exactly the same docility as the female paupers. There are three male writers that express a respectful attitude: one says 'so I Conclude and remaine your moste homble and obediant Servent' (DB/BK/4: 15+), and the second 'Your early answer to this will much Oblig your Most H^{b} Jn^{o} Hammett' (HU/BR/3: +). The third, Mr. Charles Richard Soundy who was a frequent applicant for relief together with his wife Frances, uses eloquent language as he says 'Gentlemen your parishners will in duty bound ever pray' (BK/PA/18: 28). In the other letters, the final tone is more severe. Some do not use any 'polite' final salutation but either threaten to fall on the host parish (NG/MA/1), to give the wife and children to the overseer of the host parish (HE/EA/1: 3+), to 'come down', i.e. return or send the family to the home parish on the parish's expense (OX/CH/1: 4, DO/WM/3) or to let the wife starve (OX/CL/3: 3). My impression is that the vocabulary used by men seems to be much more determined and forceful in contrast with the word choice of the female paupers, but this would of course need further investigation. It is not known, however, whether the male paupers who expressed threats in the final salutations were actually granted relief or not. As already discussed, women were supposed to show a subdued attitude when being in contact with men because of their gender, for example when writing to overseers who could grant them relief. As S. King argues, '[w]omen wrote all the time in a context full of situational limitation. When writing, they were manipulating a primarily male discourse aimed at men in power' (S. King 2019: 291). Showing a docile attitude and even asking for the overseer's advice could perhaps strengthen a woman's chances of being granted more or regular relief. Even though female paupers also threatened 'to come home', they seemed to 'bake in' this kind of intimidation into the actual narration of their distress, which reduces the impression of an actual threat in the final salutations of the letter. Here are some examples of how women presented their threats to come home, sometimes in the passive voice, sometimes by expressing gratitude through for example the emphatic 'do':

- ‘if you don’t choose to send us more than you do we must come home from you humble servants William and Sarah Docwra Studham Bedfordshire I return you my sincere thanks for what I do receive from you’ (CA/RO/1:28)

- ‘i should be glad if you will send me a few lines to let me know whether you will allow me a little here till i can do for my self or whether we must come home so if i do not hear from you i shall be with you in 14 days time for i cannot live no longer without bread’ (CA/SW/S:10)

- ‘I am obliged therefore to trouble you with these lines to inform you that unless you order its continuance as usual I must be removed immediately waiting your answer by return of post I remain your most humble and dependent servant’ (DU/BC/3)

There is nevertheless one letter from a female pauper that stands out in contrast to the other ones in my subcorpus. It is a letter in a less experienced hand, but most likely autographical, sent by the applicant Sarah Harper in Herefordshire (HE/BR/1: 9+) who demonstrates that she is aware of her children’s rights to relief. The hand-writing is confident, the letters are well-formed and there is no trace of phonetic writing, the language being standard English. The letter is very short but forceful. Even though there is no punctuation, the narration is straightforward and the message is clear. This might indicate two things: either Sarah Harper had had enough schooling to be able to write the letter herself, or somebody else helped her to write it. Sarah immediately complains that she is not receiving her due for the children and goes on to threaten that she will bring them to the home parish with a pass. Apparently, the parish had already informed her that the relief would be stopped, and yet Sarah Harper insists on her children’s right to it. There are no apologetic words or phrases in her letter, and the tone is harsh. She opens her letter with: ‘Sir I am informed that you have stopped my pay for the children and if you have you may depend on it that i shall be obliged to bring them over with a pass and then that will be expensive as you cannot deny them of the parish’. This demanding tone is unusual for a working-class woman, compared to the other letters. One may wonder if this is due to gender or to the awareness of a pauper’s right to out-parish relief.

Lexical differences (or similarities)

The LALP pauper letters from my subcorpus, both those coming from women and men, show the same preoccupations, i.e. the situation of financial distress and the application for relief or any other help from the parish such as clothes, shoes, and also help to find an apprenticeship for a child. This situation necessitates verbal pledges concerning that which lies closest to the applicants, so it is no surprise that the words used concern the children, old age, illness, unemployment (or rather no work), the lack of bread, starvation, need for clothing, the high cost of necessities, the low wages that cannot support a whole family and real distress but not wanting to be troublesome. The vocabulary used by both genders does not seem to vary greatly except, of course, that women speak of their husbands and children while men speak about their wives and children. Some letters come from older paupers that cannot work anymore, and they often use a simple vocabulary concerning their advanced age and their illnesses, and sometimes they also mention their children that cannot support them.

As S. King (2019) argues, 'the tactics adopted by those who negotiated with their settlement parishes on paper might differ from the tactics of those who could approach the overseer in person' (S. King 2019: 43). Showing exterior signs of poverty as well as a capacity to persuade in a face-to-face situation must have been convincing proof to obtain relief. Direct and immediate access to the reaction of the interlocutors through body language was also certainly an advantage compared to attempts of verbal persuasion through a written letter. The immediate response of the overseer could probably be anticipated by observing his gestures and mimics, and a less successful introduction could be turned into a fruitful conclusion on the spot in an encounter in the home parish. These conditions were not accessible to the paupers who lived far from home and had to write for relief. They were therefore compelled to develop certain epistolary strategies to obtain relief. These applicants were thus forced 'to spend more time conveying their trustworthiness and deservingness than would someone applying in person, who could be viewed easily by ratepayers and officials' (S. King 2019: 43). If there are any differences between the language of female and male paupers, these differences might then not lie in any specific gendered vocabulary but more in style, writing conventions and societal expectations. Palander-Collin, Nevala & Sairio (2013) argue that '[t]he language of

letters transmits and creates belonging to a certain group, which can be based on a relatively stable group membership of social rank or gender or on other factors defining "us and others"' (Palander-Collin, Nevala & Sairio 2013: 291). Considering this question of belonging and identity, we saw in the previous section on the address form and final salutations that the tone of the letter may vary according to the writer and her or his position in society as well as the expectations of the addressee.

It seems that most pauper writers tried to elevate their language style by using standard phrases for epistolary exchange as well as some non-native words deriving from Latin and French, for example 'necessaries' and 'liberty'. But even Germanic words such as 'know', 'children' and the days of the week caused some hesitation when it came to standard spelling (Smitterberg 2012: 955–957). Consequently, having not received much schooling and probably no guidance in letter writing, paupers did not know well how to spell the words they used. They might have picked up certain words and phrases from other paupers or friends from the lower and middling classes who had themselves some experience of epistolary exchanges, but this is not certain. Neither is it very clear if paupers had access to manuals or guides on how to write letters as the only place they probably came into contact with printed material was in a school environment (Laitinen & Auer 2014: 194–195). As already mentioned, paupers probably did not have access to any specific model when writing their letters even though they might have learnt and copied words and phrases from printed texts or other pauper petitions (S. King 2019: 142–144).

Since girls were less likely to attend day schools and their curriculum was based on a 'family ideology that stressed the moral importance for society of girls being educated for a future state of wifehood and motherhood', working-class parents 'might have considered the education of their daughters as less important than that of their sons' (Purvis 1989: 76). Would it then be logical that male paupers used more loan words than female paupers? The vocabulary linked with the relief requests and epistolary exchanges seems to consist more of words with a Latin or French origin. Through a qualitative analysis, it is nevertheless difficult to assert that male paupers were more acquainted with foreign words, but the quantitative approach might reveal certain clues to this. What could be said about the vocabulary used by women is that they seem to be more at ease with standard phrases in relation to the importance of the epistolary frame in the final salutations

of their relief requests. This is also something that can be either confirmed or refuted by the quantitative approach (see 4.3).

Apart from the general distress and poverty invoked by female paupers and the expressions of their subordinate social position, there is another concern expressed by some women and that is conveyed through the lexis: the acquisition of a marriage certificate. This certificate represented a formal proof of which parish they belonged to and where they were entitled to apply for relief (DB/TI2:12r; OX/CL/1:1). As a consequence, poor women might have asked the parish for a marriage certificate more often than men, as this could have entitled them to obtain relief. Male paupers were not directly concerned by this, as their status did not change when they married, contrary to the women who acquired their husband's home parish.

Punctuation

Today, the use of punctuation seems normal to us, but in earlier centuries, this was not so. The English style of punctuation is based on the Greek and Latin punctuation of the classical period[1]. Punctuation was originally a help in elocution, and it is only with the invention of the printing press that the syntactic punctuation system started developing (Salmon 1988). As can be expected, there were periods when punctuation was widely used, for example during the 18th century. It was only in 1906 that a concise British punctuation was suggested by Henry Watson Fowler and Francis George Fowler in *The King's English* and rendered popular.

As mentioned earlier, there is generally no punctuation, or only very little, in pauper letters. As for the amount of punctuation marks in the twenty letters of my subcorpus, it does not differ from the other LALP pauper letters. There is a general lack of commas and full stops all through the letters, except for the abbreviations in dates and months (for example 'Dec^{r}', with a dot below the 'r') and other words that were commonly abbreviated, for example 'humble', 'servant' and 'obedient'. Dots are frequent in such abbreviations and seem to have been a convention known by most paupers writing relief letters. Since punctuation was not standardised and did not seem to be of much importance, it might not have been taught in schools attended by paupers, just as writing

[1] for a brief history of punctuation, see https://www.britannica.com/topic/punctuation

was not systematically taught to all children attending school (Smitterberg 2012; Auer & Fairman 2013). Without punctuation, the ideas and logic of the applicants melt into the narration, and this can seem bewildering for modern readers. The overseers received many such letters and were probably used to decoding them, as 'common experiences and structures generated a shared linguistic register for both applicants and recipients' (S. King 2019: 117). Overseers also presumably knew the applicants personally in many cases, which might have made it easier to understand the text of the letter and to cope with the spelling variation as well as the lack of punctuation.

Traces of standard writing

The extensive spelling variation, phonetic writing and lack of punctuation gives the impression that pauper letters do not contain any standard language at all. This impression is, however, not totally justified. In many aspects, and probably unintentionally, paupers adhere to the grammar rules that were drawn up by the 18th-century grammarians, probably relying on their own speech and writing practices, as '[l]earning to spell included learning to spell in the modern sense and dividing polysyllabic words into syllables' (Fairman 2007: 198). As mentioned earlier, paupers seemed to be aware of some of the letter writing conventions, and they made efforts to adopt the standard phrases and words of the epistolary frame as much as possible. The initial and final salutations in the pauper letters seem to correspond to standard English as used by educated writers, in spite of spelling variation. Relief applicants made frequent use of standard words and phrases such as 'Gentlemen' at the beginning and 'I am your humble servant' at the end of their letters. Politeness, a frequent element in standard epistolary language and associated with the polite language of the elite, was also a common strategy to entice overseers to provide help and to awake pity in other recipients, such as the vestry clerks and wealthy parish landowners (S. King 2019: 177, 192). All these elements are, however, tinted with oral features and variation in spelling and do therefore not entirely correspond to the English language promoted by for example 18th-century grammarians and writers that recommended a standard spoken and written language through dictionaries, grammars, manuals and guides.

Another language feature that does not correspond to standard writing is the systematic use of the coordinating conjunction 'and' to connect ideas and phrases (see Fig. 4.1 and Fig. 4.2). Culpeper & Kytö (2010) argue that

> the grammar of spoken conversation proceeds incrementally, chunk-by-chunk, and the chunks are often held together by coordinators, such as AND, performing a discourse function of some kind. One of the particular functions of AND here is that of 'narrative AND', used to structure a narrative [...]. (Culpeper & Kytö 2010: 98)

As punctuation was not used, there was a need to make the narration advance in order to disclose the arguments that could lead to the granting of relief, and the easiest way to do so seems to have been the use of 'and'. This word is used 28,364 times in the Bible (King James' version from 1611)[2], so perhaps paupers were influenced by this text that they certainly were familiar with (cf. Laitinen & Auer 2014). Since the writing of paupers is speech-like, it is not entirely surprising to find this feature in their letters. The quantitative analysis confirms that this coordinating conjunction is frequent in comparison to other words (see 4.3).

The standard use of capital and small letters did not seem to be known by relief applicants, neither female nor male paupers, and this is also an element that diverges from standard English. As mentioned above, capital letters could be used anywhere in a word, especially at the beginning of a word but also inside the word. As there is an evident lack of punctuation, it is not easy to see when a sentence ends and another starts, which makes it difficult to use the standard rules of capitalisation. Some paupers seem to use capital letters for the words that they consider important while others use uppercase letters randomly or at the beginning of a line. This is, however, not systematic. This would certainly be an interesting issue for further research, as a pattern might appear in a corpus such as the LALP. The case of proper names is also interesting, not necessarily only from a gender point of view, since the writing of both female and male paupers show this tendency, although I have not methodically investigated this. Proper names are not systematically capitalised in pauper letters, and there are many examples of variation in names, even in the same letter. In the letter from John Fogg on behalf of his parents

[2]https://www.artbible.info/concordance/a.html

(DB/BK/4: 15+), the letter writer gives proof of the confusion of whether to use capital letter or not in names. He switches between 'Miss mills', 'miss mills', 'Miss Mills' and 'John mills', which can seem surprising.

One recurrent feature is the variation of the personal pronoun 'I'. Usually, the standard writing is 'I', with a capital letter, and this also seems to have been the case in the 18th century. In most pauper letters of my subcorpus, the applicants use the capitalised version (for example DU/BC/3), while a few do not seem to know this convention and write consistently 'i', the dotted version (for example CA/SW/5: 10). In some letters, both capital and small letter of the personal pronoun can be found (DB/BS/6: 1, CA/TR/2: 31). Sarah Leeland (CA/SW/5: 10) surprisingly employs one first time the capital 'I' while in the rest of her relief request, she uses the small letter. In his plea sent from Carmarthen (HE/EA/1: 3+), Richard Jones also mixes both versions in an erratic manner (HE/EA/1: 3+), as does Elizzabeth Wootton in her request (CA/TR/2: 31).

Phonetic writing, lack of punctuation and inconsistent capitalisation are to be found in all pauper letters of my subcorpus, and standard writing as promoted in the 1700s is rare. Standard language may nevertheless appear when an official or a schooled writer has helped or written the letter for the pauper. The letter by Sarah Harper is the only letter in my subcorpus that shows standard English, and it is likely that she wrote it herself. Even though it contains no punctuation, this letter shows no traces of oral features at all. It does, however, contain a mix of both capital and small 'I', although the language seems very standardised. In my opinion, the overall impression of the pauper letters is, however, that paupers tried to write standard phrases and standard language in spite of their lack of schooling.

Oral features (phonetic writing)

In the section on the main characteristics of the LALP pauper letters, I described certain features that are common in pauper relief requests, among others phonetic writing. My subcorpus does not differ from these general characteristics: The presence of oral features is striking in all the letters from both female and male paupers. There are, however, only a few examples of h-dropping in my subcorpus, even though this feature could be expected in all pauper letters. 'I ope' can be found in few letters (HE/EA/1: 3+, DB/BK/4:

15+), which can seem somewhat surprising. Or perhaps not: When writing, paupers might have preferred inserting a few extra aitches to be certain their writing was correct instead of dropping them. There are quite a few hypercorrections of this sort in my sub-corpus. Only to mention a few, there are 'I ham' which is quite frequent (HE/EA/1: 3+, OX/CL/3, 3+, OX/CH/1: 4, DO/WM/3, LA(B)/MI/4, CA/SW/5: 10, DU/BC/5), as well as any variation of the words 'is' and 'ill', for example 'my little children his very hill' (HE/EA/1: 3+). In his long explanation about the situation of his parents (DB/BK/4: 15+), John Fogg uses both h-dropping and hypercorrection in the same sentence which seems to give a good indication of his language variant as he writes 'she Will be at the trouble of Paying him has the Parish as behaved so bad'. In this sentence, there is both hypercorrection ('has' instead of 'as') and h-dropping ('as' instead of 'has').

Standard	**Var1**	**Var2**	**Var3**	**Var4**	**Var5**
overseer	oveseer	overcere	overseere	oveerserr	oversear
relief	reliefe	releife	releif	releafe	–
necessaries	neſasarys	neceſsaries	nessereys	nessaries	–
oblige(d)	obledged	ableg	oblight	oblig	–
children	childern	chirldren	cheldren	childran	cheldron
week	weak	weeke	–	–	–
write	rite	raight	–	–	–
wrote	wrot	rote	worought	–	–
please	pleas	plase	–	–	–

Table 4.1: Spelling variation in the LALP corpus.

The letter sent by William King (DB/BS/6: 1) is an interesting example of phonetic writing as well as that of Richard Jones (HE/EA/1: 3+) that I mentioned earlier. William King's language variant stands out well in certain words and phrases – instead of

'I should have been' he writes 'I should of been', 'forther' instead of 'further' and 'ourder' for 'order'. In the letter BK/PA/18: 28 from Charles Richard Soundy in Berkshire, we find 'git' for 'get', and James Dacombe from Dorset (DO/WM/3) writes 'Come Doon' twice in his letter instead of 'come down'. In Sarah Leeland's letter (CA/SW/5: 10) from Derby, we can see that she uses 'wather' for 'whether', 'longor' instead of 'longer' and 'icold' for 'I could', which might give an indication of the pronunciation in the county she was from. An interesting fact is that Sarah Leeland adopts 'icold' in the beginning of the letter, and a few lines further down reverts to the two-word and more standard version 'i Could'.

These oral features are all interesting, because they reveal speech-like written language. Phonetic writing indicates how the labouring classes probably would have pronounced the words they were writing, since they did not necessarily have access to written standard language. As spelling variation probably due to phonetic writing is frequent in pauper letters, it is impossible to carry out a quantitative analysis of the letters as such. In order to proceed to such an analysis, it is important to standardise the language, process that I used on my subcorpus with VARD2 and that I will explain in the next section.

Table 4.1 gives a few examples of spelling variation from the counties Hertfordshire, Leicestershire, Suffolk, Warwickshire and Wiltshire, spelling variation observed during the transcription of pauper letters. These examples are not meant to serve as a comparison between counties or between paupers, but only as an illustration of the existing inter-writer variety in the counties mentioned above.

4.2 Spelling normalisation with VARD2

Standardising the language of pauper relief requests is no easy task, as there is a great variety in spelling, and to this day, there exists no specific computer program for the standardisation of Late Modern English working-class language. I therefore used the VARD2 program to carry out the spelling normalisation of the letters in my subcorpus, hoping that the standardisation process would be accelerated compared to manual standardisation. VARD2 was created for the use of another type of corpus, dating from the Early Modern English period (Baron & Rayson 2008). The process of spelling normalisation of the pauper letters in my subcorpus with the VARD2 program proved to be laborious,

necessitating much manual intervention and corrections in spite of the program being efficient in spelling normalisation.

After having cleaned the plain text versions of overseers' comments, their addresses, tags, transcription annotations and archive comments, I ran the letters in the VARD2 program to obtain modern standard spelling. Some problems were encountered with when normalising the texts, as I had to decide exactly what to standardise. Was the standardisation of phonetic spelling and spelling variation sufficient, or would it be necessary to correct the punctuation and capitalisation? And what about incorrect verb forms? Phonetic spelling and spelling variation did not seem to pose a problem as VARD2 made many suggestions and kept learning new variations as I went along. Even though the important spelling variation caused a great deal of manual work, the transformation went rather smoothly. Punctuation and upper- and lowercase letters, however, turned out to be a more complex problem. As modern standard English comprises punctuation and capital letters in a codified manner, I had to decide whether to add punctuation in the standardised versions of the pauper letters or not. If my choice fell on adding punctuation, this would automatically entail adding the right capital letters in the adequate places. These questions proved to be quite a challenge. As already mentioned, the narration and ideas of paupers resemble a flow of thoughts, and it is difficult to determine where one sentence ends and another one starts. This process seemed to me like a translation operation as it became my responsibility to translate the intentions and thoughts of the paupers. I am not convinced this kind of 'translation' would guarantee a faithful result and respect the original letters of the paupers. That is why I decided not to add punctuation but only keep the one already present in the originals.

I then still needed to decide what to do with the erratic capitalisation. Some paupers used capital letters at the beginning of words and sometimes in the middle, not necessarily in a systematic manner. Even the personal pronoun 'I' could be upper- and lowercase, even in the same letter. Finally, I set up a few rules and followed these as closely as possible:

- The original punctuation was kept, if there was any. A capital letter at the beginning of the next sentence was not added, in case it was written with a small letter. No extra punctuation was added in the letters. Certain lines, hyphens or other

marks without any specific apparent meaning or function were deleted. Underlining was also left out and not mentioned in the standardised versions of the letters.

- Capitalised letters inside words and sentences were consistently translated as lowercase, with a few exceptions. In the standardised version, uppercase letters were kept only if there was a reason for capitalisation. Such cases are the names of months (November, March, etc.) that are generally spelled with uppercase today, as well as the words 'Gentlemen', 'Sir', 'Overseer' and variations of these, comprising plurals. These words seem to be capitalised for reasons of politeness or respect towards the authority. Place and person names are kept as they were initially written: if they were written with small initial letter, they remain so. As examples of this, I can mention 'greenwich' (initially 'grenwich') instead of the standard 'Greenwich' (BF/SH/2) and 'miss mills' instead of 'Miss Mills' (DB/BK/4: 15+).

- The personal pronoun 'I' was not standardised to uppercase. In cases where 'I' was spelled with a small letter, i.e. 'i', this graph was maintained. The reason for this is that the capitalised version of the personal pronoun may indicate an awareness of standard English, and it could be interesting to see whether there is a difference in the use of these two variants between female and male paupers.

- Certain verb forms have been altered, for example the 3rd personal pronoun -s has been added when absent, as well as the past participle in irregular verbs when not corresponding to standard English. The form 'do' and the negation 'dont' in 3rd person singular have not been changed to 'does' and 'doesn't' as these forms might be important indicators of differences in the language between female and male paupers.

- In order to standardise the language of the pauper letters, I chose to include the date of the letter whenever that was present, as well as the applicant's address that generally appears at the end of the letter. The address of the recipient is, however, not part of the standardisation process.

- Some words or letters were difficult to read because of holes, tears, blots and other damages to the letter paper or because of self-corrections (crossed out, inked out). The meaning of such words can sometimes be deduced by the context, and I there-

fore decided to keep the whole word in spite of the uncertainty it represented. If the damage was too important, the word was left out.

- Some longer words would have needed hyphenation as they started at the end of a line and continued on the following. Hyphens were nevertheless not added and thus one word might be translated as two in the quantitative analysis.

Apart from the issues that I encountered and have discussed above, another problem arose. I felt I was not entirely acquainted with the different options in the VARD2 tool in order to be comfortable when navigating the program. This prevented me from taking full advantage of the options that were present. Some training in the use of this program would be high on my wish list in case I would need to use VARD2 again.

When using VARD2, I observed that the word variants are numerous because of the important amount of spelling variation, and that the first automatic suggestions provided by the program are far from adequate. It was therefore necessary to manually insert the corresponding words to most of the variants, process that was time-consuming. There were some words that are unknown today or that had not been understood by the transcriber. These remained as they were spelt, even though their meaning was not elucidated during transcription and could thus not be standardised.

As a conclusion, I would like to highlight three main points. First, when standardising Late Modern English spelling, especially language from pauper letters presenting phonetic spelling, it is essential to have a very good working knowledge of the computational tool, be it VARD2 or any other computation appliance. If this is not the case, collaboration with a computational linguist or with a person well acquainted with the chosen tool is recommended. Second, the standardisation needs to be approached as a translation from one language to another. The phonetic spelling gives many clues to the language of the paupers, and when standardising it, many important linguistic aspects might be lost. It is therefore very important to accommodate the standardisation process to the research question. An automatic standardisation might prove fatal to certain linguistic aspects and it is therefore crucial that scholars and scientists know the source texts and their characteristics before making any standardisation attempt. Finally, I would like to emphasise that even for a small subcorpus, the spelling normalisation required extensive work and laborious checking.

4.3 Quantitative analysis

To be able to proceed to a quantitative analysis, it is important to have texts in standard English to analyse. As the LALP pauper letters present much phonetic spelling and spelling variation, it would be complicated to perform a quantitative analysis on such material. A standardised version of my subcorpus with the VARD2 program thus proved to be necessary in order to produce a reliable quantitative analysis. The texts that I used for this analysis are therefore the standardised text versions that I generated with the program VARD2.

In order to proceed to the analysis, I chose the concordance program AntConc that is freely available online[3]. I have already used AntConc on other texts and am somewhat familiar with it although not an expert. Another possibility would have been the program LancsBox 6.0 provided by Lancaster University, but I decided not to operate my analysis with this program because on exploration, it seems more complex than AntConc. A more complex tool requires a good user command, which is not the case here in this situation.

I carried out several analyses with AntConc, looking at word lists and frequencies, grammar (such as *that*-clauses) and vocabulary. The length of the letters was also an interesting point that I wish to discuss below.

Word count

First of all, an interesting fact is the length of the letters[4]. Letters written by both female and male paupers can be either long or short, but on the whole, men use more words in their letters than women. The average word quantity for letters written by men is just over 209 words and only approximately 160 for women. The longest letter written by a male pauper was 404 words long while the longest one written by a female pauper contained 269 words (see Table 4.2). This of course raises several questions: why did male paupers in general write longer letters than female paupers? Could it be that men – even paupers – had had more schooling in writing than women and were more capable of

[3] see https://www.laurenceanthony.net/software/antconc/

[4] The word count is based on my 'cleaned' text versions and not on the information given in the LALP metadata of each letter

Letter nr	Women (words)	Letter nr	Men (words)
CA/RO/1: 28	204	DB/BK/4: 15+	279
CA/SW/5: 10	176	DB/BS/6: 1	404
CA/TR/2: 31	126	NG/MA/1	189
DU/BC/3	88	HE/EA/1: 3+	282
DU/BC/5	190	HU/BR/3: 8+	125
OX/CL/1: 1	170	OX/CL/3: 3	80
BF/SH/2	269	OX/CH/1: 4	190
DB/BK/1: 4	89	BK/PA/18: 28	235
DB/TI/2: 12r	183	DO/WM/3	107
HE/BR/1: 9+	102	LA(B)/MI/4	200
Total	1597	Total	2091
Mean	**159.7**	**Mean**	**209.1**

Table 4.2: Word Count

writing letters? Educational history could certainly give an answer to this question. Did female paupers use less words in their letters because their vocabulary was smaller due to a lack of education? Or perhaps female paupers did not need as many words as male paupers in order to convince the overseers of their distress? In the 18th century, women seem to have been much exposed to the risks of widowhood and poverty as they were not trained in a profession and husbands could die early, for example in wars or in work accidents (Hill 2013: 240–241).

Being considered as more fragile than men and needing protection from male members of society, women in the 18th century might also have benefited from this inferior social position in the case of relief. This situation might influence on the quantity of words needed to sound convincing in the relief request. A widowed pauper woman with children apparently needed more protection than a widowed man in the same situation. The woman might have found it hard to marry again, bringing the children of another man into the second marriage (Hill 2013: 241), while the man was still expected to continue working and perhaps simply remarry a (younger) woman to care for the home and existing children. The fact that the letters of male paupers in my subcorpus are in general longer than those of female paupers is, however, an interesting fact. If this analysis was made on the whole LALP corpus, would the result be the same? Exploring the entire LALP corpus could probably give us more exhaustive answers than I can give here.

Word lists

The fact that letters written by pauper women are generally shorter than letters by pauper men raises further questions – is it because women had a smaller vocabulary, or a different one, from men? Or did women use a more Anglo-Saxon lexis than men? In order to examine quantitatively the language found in the pauper letters, I started with making word lists with AntConc. When comparing the word based on the letters written by male paupers with those by the female paupers, several interesting features stood out. The capital 'I' was found 76 times in the letters written by pauper women, while small 'i' appeared 19 times. The percentage of the standard capitalised 'I' found in the letters written by women is thus 80%, and 20% for the small 'i', which could indicate that female paupers were somewhat unsure of how to write this personal pronoun in standard language. In the relief requests by male paupers, I only found 12 occurrences of small 'i' for

102 instances of the standard capital 'I'. The percentage for males is thus 10,52%, which is considerably less than for the group of female paupers.

Rank	Freq	Word
1	76	I
2	65	to
3	65	you
4	38	and
5	38	the
6	34	me
7	29	as
8	29	for
9	28	of
10	27	my
11	24	have
12	23	a
13	19	am
14	19	i
15	19	it
16	19	will
17	17	not
18	16	be
19	16	if
20	14	do

Figure 4.1: Word list female paupers

Put in relation to the average number of words that male paupers wrote in their letters, what could these results indicate? On the average, male paupers wrote longer letters than female paupers and used a capital 'I' more often. Could this indicate that male paupers were more acquainted with standard English than females? Or could this be connected with the importance letter writers put on some words and less on others by capitalising them? It is possible that some paupers proceeded in this way, capitalising such words they found relevant in the context. Perhaps female paupers might have known that the 'I' was capitalised in standard English but did not want to appear pretentious in the eyes of the overseer by making themselves look important through a capitalised 'I'? It would be interesting to search the whole LALP corpus for any specific signs indicating that women writers used the small 'i' more often than men, since my subcorpus shows this tendency.

When working on the word lists, I noticed that for the letters written by male paupers, there were 538 word types and 2055 word tokens, while for women, the same categories landed on 463 (types) respectively 1589 (tokens). Could this mean that female paupers had a narrower vocabulary than male paupers? If women used a more restricted vocabulary, what could be the reason for this? Social expectations, or simply the fact that girls had different curricula than boys at school (see Purvis 1989)? As already discussed, girls were encouraged to learn sewing and other subjects in connection with the sphere of the home while boys were not. Neither were women to speak out but they were to remain outwardly humble and silent, or at least avoid gossiping and unnecessary talk. Could the

differences in types and tokens stem from others reasons, and if so, which ones? This is also a point that would deserve further attention and exploration.

There were also other singularities that drew my eye when viewing the word lists and comparing them, and I will briefly mention these here. The first fifteen most frequent words corresponded fairly well between female and male paupers, with one or two exceptions ('that' and 'i'). Female paupers were, however, more prone to use 'Sir' (23rd position and 12 occurrences with capital letter) than male paupers (30th position and 11 occurrences). 'Sir' is usually used as the address form and was positioned straight after the date but could also appear in the body of the letter. The word 'Gentlemen' is also more frequently employed by female paupers (frequency position 27 with 10 occurrences) compared to males (frequency position 117 with only 3 occurrences).

Rank	Freq	Word
1	102	I
2	86	to
3	64	and
4	64	you
5	63	the
6	42	for
7	37	in
8	35	me
9	33	have
10	33	of
11	32	that
12	30	will
13	29	it
14	26	as
15	25	my
16	25	not
17	25	so
18	24	am
19	23	a
20	23	is

Figure 4.2: Word list male paupers

Women seem to use 'Sir' and 'Gentlemen' as a vocative (cf. Fairman 2000: 71) to address themselves in a more diplomatic manner to the male recipients than male paupers do. At least women seem to plead with the overseers in a different way than men. The role of the addressee seems therefore important: who will receive the letter, and who will read it? What social position does this person occupy? Historically, women 'are generally granted less status and power than men', and '[b]y using prestige language forms, women wish to assert their authority and position and to gain respect' (Nevalainen & Raumolin-Brunberg 2017: 111). It could thus be argued that women's respectful and 'polite' attitude is a direct consequence of the social stratification where women, especially women from the lower social classes, were to be found at the bottom of the social hierarchy (cf. the models of social stratification in Nevalainen & Raumolin-Brunberg 2017: 136). Even then,

they probably knew the importance of fulfilling social expectations efficiently through efficient 'female' language.

Another interesting result is that paupers used a great amount of monosyllabic words when they wrote their relief requests. The word lists for both female and male paupers reveal that words with two or more syllables are less frequent and appear later on in the list. For female paupers, the first word with three syllables is 'Gentlemen' (26th position), while for males, the same word ranks 100th. Other words consisting of three syllables are frequently place names ('battersea', 'carmarthen', 'lymington') as well as weekdays and months. In the word lists, two-syllable words start appearing around the 40th position, with 'children', 'distress' and 'shillings' for men and 'parish', 'humble' and 'answer' for women. The word 'consideration' appears only later on in the word frequency list (94th position for males and 95th for females) as well as 'overseer' (males 321st position and females 275th). And yet, even though there is a marked tendency towards the use of Anglo-Saxon monosyllabic words in pauper letters, it is clear that paupers also were acquainted with Latinate words (Fairman 2000: 72). Interestingly enough, it seems as if paupers had a tendency to create several monosyllabic words out of words with more syllables. This is the case with words starting with 'a', for example 'a quaint' (two one-syllable words) instead of the two-syllable word 'acquaint'.

If we take a closer look at the two word lists, the resemblance is striking – both women and men use about the same words in their relief requests. As already discussed, the personal pronoun 'I' is very frequent for both women and men as well as the coordinating conjunction 'and' (see Culpeper & Kytö (2010: 173–175) for a more in-depth analysis of this in Early Modern English dialogues). On the other hand, there is one word that is more frequent in the word list of male paupers, and that word is 'that' which I will discuss in the next subsection.

That-clauses

I would now like to draw the attention to the word 'that', appearing in the frequency word list. There is a great difference in the use of the word 'that' between female and male paupers. Male paupers used it 32 times throughout the ten letters and its frequency was placed in the 11th position. For female paupers, however, 'that' ranges itself as 22nd in the

list, with only 13 occurrences. Grammatically, the word 'that' can signify many things – is this the pronoun, the determiner, the relative pronoun, does it introduce a *that*-clause, or is it used as an intensifier? In order to discover the role of this word in the letters, I checked it with the concordance tool. The result is very interesting: the concordance chart shows that female paupers mainly used the word 'that' to introduce *that*-clauses (see Fig. 4.3), often after the verbs 'inform' and 'hope'. Only twice do female applicants use 'that' as a pronoun. The chart of the male paupers' use of the word 'that' shows a more varied result (see Fig. 4.4). This word was used as a pronoun or determiner at least five times by male applicants. The general use is also for *that*-clauses, but compared to the female paupers, the male paupers use more varied verbs linked to 'that' ('hope', 'inform', 'acquaint', 'tell', 'think', 'say'). Nouns and adjectives were also employed to a certain degree to introduce the *that*-clause. Male paupers therefore seem to make a more intensive and diverse use of 'that' for *that*-clauses than female paupers. The reason for this is not clear: Could this be associated with the level of schooling, or perhaps male paupers being more acquainted with letter writing?

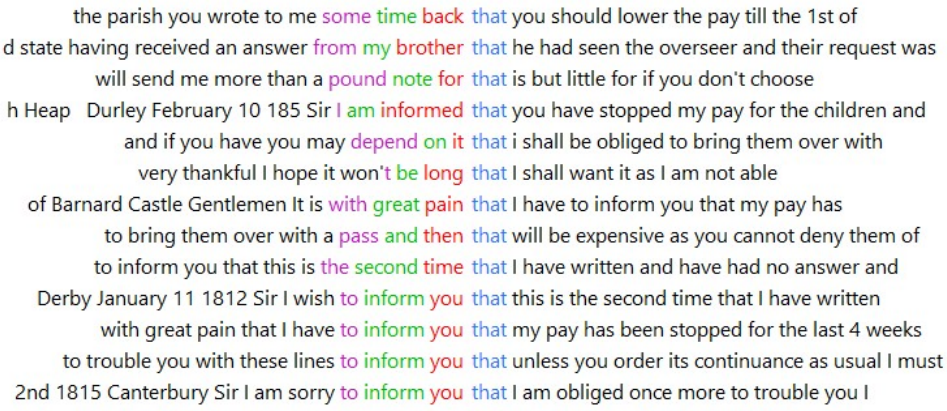

Figure 4.3: Female paupers: *that*-clauses

Vocabulary

From the word lists, it is quite obvious that paupers preferred Anglo-Saxon monosyllabic words but were acquainted with foreign words stemming from Latin and French. The question of the vocabulary of paupers' relief requests is a topic that would deserve a more serious analysis than I can make in this *mémoire*. It seems that many words associated with letter writing and relief applications are not Anglo-Saxon but rather Latinate: 'obedient',

out of my wages I hope in the Lord that Gentleman and Overseers will not neglect to send
us, for times is so dreadful bad at Nottingham that there is very little work to be got I
it does pay my milk and foam charges and that is taken of whether I have anything or not
's place on Monday next 18th of May as that is the day appointed for us to meet to
you word that I am in such great distress that unless you send me some relief on next Saturday,
rest in the night so that I cannot do that work as I should do and then my work
for 2 shillings a day and great favour to get that there is 2 shillings for fire and four shillings
not at Kingston but my wife is I hope that you will not fail sending at the time as
bold in addressing a few lines to you hoping that you will not be offended at me so doing
only my wife and the children are so ill that I cannot therefore unless you send me some relief
I am very sorry that I have to inform that my wife she is still very ill and has
for fire and four shillings a week for lodgings that is 6 shillings out of my wages I hope in
in this case I am brought down so low that I now am quite tired of my life in
went with my father and mr Barnett informed me that they did not make a practice for to advance
I am but a weekly tenant he tells me that he will take our fire goods and turn us
would not grant him any relief but told me that they would write unto you and if you would
I was in arrears at the shop for meat that they have refused me anymore the the day 12 month
a shirt to put on I do not mind that I if you do not send I must come
him what Miss Mills told me and he said that i had better write and inform you of it
I have since seen Miss mills and she says that if you will be so good as to send
wash and lose my rest in the night so that I cannot do that work as I should do
Gentlemen and Overseers of Baslow I am very sorry that I have to inform that my wife she is
way of going on and I am very sorry that I do lay so heavy upon you but I
they both do think that it is something strange that you have not been nor sent to the present
Wolley and Mr Thomas Hogg they both do think that it is something strange that you have not been
at Baslow 3 weeks ago for Mr wolley he thought that it would be much better for me to come
I do not wish for to put you to that expense but I should not have put you to
you but it is not without being in want that I write again to you As I mentioned in
none so I am compelled to send you word that I am in such great distress that unless you
at me so doing it is to acquaint you that on Friday last it being the day for the
April 15 Sir) these few lines is to inform you that I went to Mr Jackson the carrier expecting some
December 26th Sir I am sorry to inform you that we are in Carmarthen we are in very great

Figure 4.4: Male paupers: *that*-clauses

'servant', 'inform', 'application', 'obligate', 'oblige(d)', 'liberty', 'consequence', 'acquaint' and 'consideration' are derived from Latin or French and are much used in relief letters. The lexicon related with relief, a system established by the elite to tackle the poorer part of the population, seems also to be closely connected with a Latin vocabulary: 'assistance', 'support', 'subsist', 'dependent', 'marriage', 'certificate', 'necessaries', 'opportunity', 'petitioner' and 'suspension' are also frequently found in relief requests.

Apart from the words already discussed above, the most frequent words are (as could be expected) verb forms ('will', 'do', 'can', 'could', 'send', 'be', 'is', 'hope', 'have'), nouns ('parish', 'children'), prepositions and infinitive mark ('to', 'of', 'if', 'for', 'at'), pronouns and possessives ('me', 'my', 'you', 'your'), determiners ('a', 'the', 'this') and other words such as 'as', 'not', 'so' and 'please'. Surprisingly enough, the word 'child' and its plural are used almost equally by both men and women, and the word 'relief' is only mentioned four times in letters written by pauper women and six times in those written by men. Paupers used other words to express the relief they hoped to be given, by calling it 'a favour', 'help', 'something', 'my pay', or simply 'it' and spelling out how much was already granted in pounds and shillings and how much was needed, for example for rent, clothing, food, or for other types of debts paupers might have contracted. Both paupers and overseers knew the cost of bread and everyday articles, and it might thus have been a better idea to be precise moneywise instead of remaining in a more unspecified request of financial help.

As a conclusion, it could be argued that paupers had an adequate vocabulary related to letter writing and relief requests and that there are no marked differences between the vocabulary used by female and male paupers. Smitterberg argues, however, that

> during the early part of the Late Modern period, knowledge of classical languages was reserved chiefly for men [...] which may have affected women's command of loanwords from Latin and Greek. As far as orthography is concerned, Görlach (2001:57) argues that women's inferior educational opportunities led to more 'vagaries of spelling' in women's 18th-century writing than in men's. (Smitterberg 2012: 956–957)

I believe the main issue was that paupers had received no or little schooling and did not know how to spell the words according to standard language rules, even though

the words were known to them. Both Anglo-Saxon and Latinate words were difficult for them to spell. No great linguistic differences seem to exist between female and male paupers as far as punctuation, vocabulary and access to standard language are concerned. The differences that might exist could be more due to socio-economical and socio-historical facts, issues that could be explored further.

I am convinced that many more interesting results could be brought forth by a more thorough analysis with AntConc or any other corpus analysis toolkit. In my opinion, all the questions that have arisen from both the qualitative and the quantitative analyses would deserve a more profound investigation by historians, linguists, computational linguists and data scientists, and I hope the conclusions expressed in this *mémoire* might be useful for future research.

4.4 Hypothesis on gendered discourse in the LALP pauper letters?

As we have seen above, there are some interesting differences in the language between female and male paupers even though they might at first glance be superficial and minimal. These differences might not be linked to gender as such but could be due to the difference in formal schooling received by both genders, social expectations, gender roles and personal capacities and interests. What S. King (2019) points out is that

> although we can certainly find women (and men) conforming to stereotypical norms in the sense of languages and signals of deference, protection, and paternalistic duty, the writing of poor women and sometimes their appearances before vestries suggest that [...] they appropriated language and gendered models just as skilfully as their middling counterparts. (S. King 2019: 292)

According to this, women from the lower social classes could also use language to their advantage in spite of their lack of schooling. As already discussed, pauper women could have used certain linguistic strategies to conform with the female role that was attributed to them in the patriarchal society of the 18th century in order to obtain certain advantages.

In the LALP corpus, many women seem to be widows with children, aged women or women whose husbands are either in the army, cripples, unemployed or taken prisoners during war. As already mentioned, the first impression I received when exploring the LAP pauper letters was that there were a great many letters coming from such women, but this is only a personal opinion and could also be a result of opportunistic sampling. This impression of females being more frequent writers of relief requests would, however, need a more profound analysis, but it still raises several questions linked to possible gendered discourse that is yet to be proved:

- As women in general were given less education than men, also in the lower social classes, how could it be possible that they were frequent letter writers? It would be important to search the whole LALP corpus to confirm whether there are more letters coming from female paupers than from male paupers. On the other hand, as has already been discussed, women were more exposed than men to misfortune and poverty since they had less education, had not access to a profession and were generally dependent on a male relative or husband. Having said this, it would be only logical that women would be more frequent applicants as they were not financially independent.

- If there are more letters written by female paupers preserved in the archives, is it possible to draw the conclusion that women paupers wrote more often than men? Or would there be any specific reason to believe that letters emanating from female paupers were more often conserved by the parishes than those written by males?

- As to the linguistic performance, were female paupers less acquainted with the formal style of the relief letter since they seem to write shorter letters? As it was the role of the husband as head of the family to deal with financial matters, were men instructed in letter writing more often than women?

- How does the supposedly lack of education influence the language used by female paupers in relief letters – they might have written shorter letters and used a more restricted vocabulary than male paupers, but does this mean that they were less capable of expressing themselves linguistically than men to achieve their goal? According to certain scholars, female paupers were just as efficient in their language

use as were women from the middling sort. But were they less convincing than men when addressing themselves in writing to the home parish and the overseer because they were women? Perhaps the granting of relief was not connected solely with linguistic performance but also with social expectations and stereotypes?

- In my subcorpus, both Anglo-Saxon words and a vocabulary derived from Latin and French were used by paupers, even though these might not have known how to spell the words. There are differences in the use of certain words and phrases, and especially the tone of the letters vary between female and male paupers. Could this perhaps be more a question of sociopragmatics and social expectations rather than of vocabulary?

- Basing myself on my subcorpus, I would like to argue that there seem to be set phrases and arguments that all paupers used when writing relief petitions to their home parish. Both women and men seem to adopt similar phrasing in their relief requests even though women seem to write shorter letters and on a different tone. It would therefore be interesting to explore the whole LALP corpus to see whether this is also valid on a larger scale. What linguistic elements give the impression of a more docile and unpretentious attitude in the language of female applicants in comparison to that of male applicants?

In the light of what I have discussed so far, my findings seem to have led me to more questions than answers, and it is therefore difficult to make any reliable hypothesis on possible gendered discourse in the LALP pauper letters. There might be consistent differences in the language of female and male paupers, but at this stage, there is no evidence to confirm this. As my subcorpus is of a reduced size, the findings based on it might not be entirely reliable and might not correspond to the results based on a larger body of texts. The results of both the qualitative and the quantitative analyses are, however, interesting and would deserve more attention than what I can give them here. A larger corpus could almost certainly reveal more information and perhaps even provide answers to certain of the questions that I have articulated here.

Chapter 5

Discussion

As discussed in the first part of my *mémoire*, corpus design is extremely important in order to obtain reliable analyses results as well as to avoid as much bias as possible. The corpus needs to be constructed to become a model of the texts it represents. It is also essential to remember that the model does not correspond to the entire original object but reflects only certain attributes of it (Piotrowski 2019b). As Piper (2017: 652) also puts it, 'models stand for something but are not to be confused with the thing itself'. Furthermore, the corpus needs to respect certain criteria, i.e. it needs to be explicit, representative, unambiguous and coherent (Biber 1993; Gries & Newman 2018; Piotrowski 2019b). The corpus thus becomes a powerful tool for the research questions that are put forward.

In order to explore any possible gendered discourse in the pauper relief letters of the LALP corpus, I created a subcorpus of twenty letters, ten written by women and ten by men, respecting the criteria for efficient corpus design. Even though my subcorpus was limited in size, i.e. the corpus was rather small with shorter texts, it can still be considered as a valid corpus (Biber 1990; Piotrowski 2019b) since it respects the criteria of a corpus being explicit, unambiguous and coherent. Although pauper letters are written evidence from the past, they mirror speech-like language of the lower social classes, and this is particularly interesting in the investigation of for example gendered discourse through history. So far, linguistic analyses and research of historical texts have to a great extent been performed on written texts emanating from the elite and the upper social classes (Fairman 2000; Auer et al. 2014; Nevalainen & Raumolin-Brunberg 2017), which means that contemporary linguistic theories might be re-modelled by new findings based

on the 'language from below' (Elspaß 2011). Pauper relief requests can therefore be considered as important and valuable contributions to put forward new research angles and results.

As described earlier in my *mémoire*, transforming historical documents into machine-readable texts necessitates several important steps: transcription and spelling normalisation need to be performed before submitting the material to more automated tasks such as corpus linguistics analyses. Most of these processes are done manually and thus require much time and efforts of all scholars involved. This work is essential for subsequent computational groundwork and should not be underestimated. When transforming my subcorpus into a machine-readable format, I was confronted with several issues that I did not necessarily expect to encounter, for example certain important choices concerning spelling standardisation and punctuation. In this case, my choices were led by my research question, the investigation of possible gendered discourse in pauper letters. Naturally, these choices could be challenged and harmonised according to the nature of the research question.

Gendered discourse is no easy topic in present-day language, and even less so for historical discourse, as it is closely linked with literacy and writing skills that were not necessarily taught in schools attended by paupers. As Smitterberg points out,

> women may have been overrepresented in the segment of the population that had acquired reading-only skills. But even fully literate women were less likely than men to be able to contribute texts to several genres: for instance, scientific and political texts by women are scarce because of women's restricted access to education and political power. (Smitterberg 2012: 955)

Women are therefore less present in historical text collections that have survived and literary production from earlier centuries. And as we simply do not have access to spontaneous spoken language from the past, it is necessary to rely on (speech-like) written sources. Pauper relief requests thus become especially important as they provide speech-like accounts, from both women and men. As the writers were poorly schooled and not always aware of standard English rules and grammar, they often wrote as they spoke – phonetically. Over the past twenty or thirty years, there has been an increasing interest in language 'from below' as well as in gendered discourse, and various studies

in the language use of women and men have been performed, but the results concerning differences in female and male speech seem somewhat ambiguous (Biber & Burges 2000). The differences between the language of women and men tend to be minimal, even though certain tendencies can be distinguished in specific situations (Labov 2001). Diachronic linguistic performance in drama between women and men has been investigated, and apparently, the results show that the addressee represents a crucial role in the dialogue exchange. Scholars found that 'female authors portrayed both female and male characters as being more involved and tentative than male authors' and that for female authors, 'the gender of the addressee was a major factor in the extent of involvement and tentativeness' (Biber & Burges 2000: 35). This comes close to what I found in the narration of paupers: the addressee seems to play an important part in the letters. The fact that female paupers wrote to male overseers might thus be of a certain importance in word choice, writing style and tone. At least, this is what my analyses seem to be hinting at.

In my investigation of gendered discourse in the twenty letters of the LALP subcorpus, I encountered certain difficulties in connection with uncertainty. When working on historical texts, it is difficult to be certain about the identities and social factors that surround the pauper letter writers: Were the paupers schooled or not? How much schooling had they had? Did they write the letters themselves or were they helped? Is a letter signed by a man truly written by him or could the writer be his wife, signing in his name? Then there is of course the 'bad data' problem concerning the representativeness of the texts. How representative are the existing LALP pauper letters compared to all relief requests ever written by paupers? It is therefore important to take this uncertainty into consideration in the final research results of the analyses.

It is also important to remember that the pauper letters are not meant to be epistolary exchanges between two members of a family or two friends. These letters were written with a specific purpose in mind in an institutional setting, viz. they are requests for financial aid sent to the parish of settlement. When in need of support, paupers who lived outside their home parish were compelled to write to the overseers if they wanted to be relieved instead of presenting an oral plea on-site. The fact that paupers had had little schooling and not much training in the art of letter writing influences greatly on the form of the letters written by paupers (Fairman 2007; Laitinen & Auer 2014). The results obtained from the qualitative analysis of my subcorpus reveal that paupers were acquainted

with the epistolary style and its set phrases such as address forms and final salutations although their spelling was not standard. There are no great differences between female and male paupers as far as this is concerned. Almost all pauper writings in my subcorpus show phonetic writing and lack of punctuation, and there are only few letters that show signs of the 'polite' standard English used by the members of the court and the gentry.

Furthermore, the quantitative analysis shows that paupers had a preference for Anglo-Saxon monosyllabic words even though they to a certain extent did use a Latinate vocabulary linked with letter writing and relief requests. Some interesting differences between letters written by women and men lie in the tone used and the docility shown to the addressee. Female paupers tend to be meeker and more subdued in tone than male paupers who present forceful threats at the end of their requests for relief. Nevertheless, it is surprising to note that women also used the same intimidation methods as men, for example to 'come down', but they succeeded in 'baking in' such threats into their narration in a way that may have seemed more acceptable to a male addressee.

In the analyses of the pauper letters of my subcorpus, I cannot find any significant major differences between female and male pauper discourse. There is no apparent concluding proof of female paupers being more prone to using standard English, as suggested by Labov (2001) for modern-day speech. It seems that the lack of schooling for both women and men had similar consequences, i.e. paupers reverted to phonetic spelling and standard phrases in connection with letter writing and relief requests. As already mentioned, men had a higher social status than women in the British society of the 18th and 19th centuries, and that is a factor to take into consideration when exploring possible historical gendered discourse. To conclude, I cannot at this stage claim that gendered discourse exists in the LALP pauper letters, only that I have found certain minor differences in the language of female and male paupers. The reasons for these differences seem to be multiple and would need more interdisciplinary exploration in order to obtain solid evidence. Basing myself on my subcorpus, however, I seem to detect certain indications pointing to disparities that could be more profound than what can be seen on the surface. These differences might be linked to historical sociopragmatics, socio-historical facts, gender stereotypes and gender discrimination in place in the 18th and 19th centuries. A thorough examination of all letters of the LALP corpus might give certain answers to what I have discussed here in my *mémoire*. This kind of research would of course require

an interdisciplinary approach, greater resources than those that have been available to me and more time to get to the bottom of the numerous questions that are still waiting for an answer.

Chapter 6

Conclusion

Throughout this *mémoire*, the importance of the corpus as a model for linguistic phenomena has been highlighted on many occasions. A carefully designed corpus and an impartial use of the material it contains are crucial elements to obtain reliable research results. As we have seen, any bias that could enter the corpus will undeniably skew the final outcome.

In order to examine the possible existence of gendered discourse in pauper letters of the LALP corpus, I created a subcorpus of pauper relief requests based on the criteria of corpus design and model-making. My subcorpus thus acted as a tool for investigating potential evidence of gendered discourse in the LALP pauper letters. It was composed of ten letters written by female paupers and ten letters by male paupers. Surprisingly enough, when selecting the pauper letters that were to enter my subcorpus, I had the impression that there were more letters written by women than by men to be found in the LALP corpus. As the LALP corpus is currently under construction, I have not been able to find any evidence of this and it remains a simple impression. The pauper letters were chosen randomly from the LALP material that presented metadata.

In the exploration of my subcorpus, I first proceeded to a qualitative analysis that indicated that language differences between female and male letter writers are minimal, except perhaps for the tone, style and the use of threats in the final salutations of the letters. Men seem to use more direct threats than women who demonstrate more verbal diplomacy, docility and humility in their requests to the overseers. As far as the lack of punctuation is concerned, there is no difference between female and male paupers. Both

genders have recourse to the same vocabulary apart from certain words that may be linked with the use of vocatives and the question of verbal diplomacy.

The quantitative analysis also yielded interesting results. After transforming the historical texts to standard modern English, process that proved to be challenging, I found that the analysis with AntConc revealed certain thought-provoking facts about the language use of paupers in their relief requests. Both women and men had a preference for Anglo-Saxon monosyllabic words even though they did use a Latinate vocabulary to a certain degree. The word count proved that in general, male paupers wrote longer letters than females, and they seem to have more word types and word tokens than females. Male paupers also made a more varied use of *that*-clauses than female paupers. The reasons for this are, however, not clear – did male paupers receive more schooling than pauper women and were thus more at ease with letter writing? Did men write more frequently than women? The difference in school curricula in the 18th century between women and men could be an explanation, but this is yet to be proven.

In my investigations, there is no immediate evidence of gendered discourse as such in the LALP pauper letters. As highlighted, there are minor linguistic differences in the relief letters between female and male paupers, but it is at this point difficult to state with certainty what they are ascribed to and if they occur in all the letters of the LALP corpus. As my subcorpus was limited, the results obtained may prove questionable and not entirely reliable, but they can still function as an incentive to further research. A thorough analysis of the entire LALP corpus would certainly provide more answers to the questions that have arisen all along my *mémoire*. To explore the LALP pauper letters in more depth, it would be necessary to approach the material from different points of view. Historical sociolinguistics, historical sociopragmatics as well as corpus linguistics would probably prove to be indispensable tools for a deeper understanding of possible gendered discourse in pauper relief requests. As we have seen, historical texts require an interdisciplinary approach, and it would be important to establish a collaboration between historical sociolinguists, linguists, computational linguists and data scientists in order to explore the LALP corpus exhaustively. This combination of resources is certain to produce stable results by combining a thorough knowledge of the historical texts and the modern computational means that are available to research today.

References

Alvarez-Mellado, Elena, María Luisa Díez-Platas, Pablo Ruiz-Fabo, Helena Bermúdez, Salvador Ros & Elena González-Blanco. (2021). TEI-friendly annotation scheme for medieval named entities: a case on a Spanish medieval corpus. *Language Resources and Evaluation* 55(2). 525–549. https://link.springer.com/article/10.1007/s10579-020-09516-2.

Andresen, Melanie, Michael Vauth & Heike Zinsmeister. (2020). Modeling ambiguity with many annotators and self-assessments of annotator certainty. In *Proceedings of the 14th Linguistic Annotation Workshop*, 48–59. Association for Computational Linguistics. https://aclanthology.org/2020.law-1.5/.

Archer, Dawn, Merja Kytö, Alistair Baron & Paul Edward Rayson. (2015). Guidelines for normalising Early Modern English corpora: decisions and justifications. *Icame Journal* 39(1). 5–24. https://doi.org/10.1515/icame-2015-0001.

Auer, Anita. (2012). Late Modern English: Standardization (Chapter 58). In Alex Bergs & Laurel Brinton (eds.), *English Historical Linguistics: an International Handbook*, 939–952. Mouton de Gruyter. https://doi.org/10.1515/9783110251593.939.

Auer, Anita. (2015). Stylistic variation. In Anita Auer, Daniel Schreier & Richard J. Watts (eds.), *Letter Writing and Language Change*, 133–155. Cambridge University Press. https://www.researchgate.net/publication/321318954_Stylistic_variation.

Auer, Anita & Tony Fairman. (2013). Letters of Artisans and the Labouring Poor (England, c. 1750–1835) (Proofs). In Paul Bennett, Martin Durrell, Silke Scheible, Richard J. Whitt, Holger Keibel, Marc Kupietz & Christian Mair (eds.), *New Methods in Historical Corpora*, 77–91. Narr Verlag (Tübingen). https://doi.org/10.1075/slcs.159.10lai.

Auer, Anita, Mikko Laitinen, Moragh Gordon & Tony Fairman. (2014). An electronic corpus of Letters of Artisans and the Labouring Poor (England, c. 1750-1835): com-

pilation principles and coding conventions. In *Recent Advances in Corpus Linguistics*, 7–29. Brill. https://bit.ly/3IjyeIG.

Auer, Anita, Catharina Peersman, Simon Pickl, Gijsbert Rutten & Rik Vosters. (2015). Historical sociolinguistics: the field and its future. *Journal of Historical Sociolinguistics* 1(1). 1–12. https://doi.org/10.1515/jhsl-2015-0001.

Auer, Anita, Daniel Schreier & Richard J. Watts (eds.). (2015). *Letter Writing and Language Change*. Cambridge University Press.

Baker, Paul. (2014). *Using Corpora to Analyze Gender*. Bloomsbury Academic.

Baron, Alistair & Paul Rayson. (2008). VARD2: a tool for dealing with spelling variation in historical corpora. In *Postgraduate Conference in Corpus Linguistics*. Lancaster University. https://eprints.lancs.ac.uk/id/eprint/41666.

Baron, Alistair & Paul Rayson. (2009). Automatic standardisation of texts containing spelling variation: how much training data do you need? In *Proceedings of the Corpus Linguistics Conference*. Lancaster University. https://eprints.lancs.ac.uk/id/eprint/42529/1/314_FullPaper.pdf.

Beal, Joan C. (2004). *English in Modern Times 1700-1945*. Arnold.

Beal, Joan C. (2016). Standardization (Chapter 18). In Merja Kytö & Päivi Pahta (eds.), *The Cambridge Handbook of English Historical Linguistics*, 301–317. Cambridge University Press.

Bergs, Alexander. (2014). The uniformitarian principle and the risk of anachronisms in language and social history. In Juan Manuel Hernández-Campoy & Juan Camilo Conde-Silvestre (eds.), *The Handbook of Historical Sociolinguistics*, 80–98. John Wiley & Sons, Ltd. https://doi.org/10.1002/9781118257227.ch5.

Biber, Douglas. (1990). Methodological issues regarding corpus-based analyses of linguistic variation. *Literary and Linguistic Computing* 5(4). 257–269. https://doi.org/10.1093/llc/5.4.257.

Biber, Douglas. (1993). Representativeness in corpus design. *Literary and Linguistic Computing* 8(4). 243–257. https://doi.org/10.1093/llc/8.4.243.

Biber, Douglas & Jená Burges. (2000). Historical change in the language use of women and men: gender differences in dramatic dialogue. *Journal of English Linguistics* 28(1). 21–37. https://doi.org/10.1177/00754240022004857.

Biber, Douglas, Susan Conrad & Randi Reppen. (1996). Corpus-based investigations of language use. *Annual Review of Applied Linguistics* 16. 115–136. https://doi.org/10.1017/S0267190500001471.

Biber, Douglas, Susan Conrad & Randi Reppen. (1998). Historical and stylistic investigations (Chapter 8). In *Corpus Linguistics. Investigating Language Structure and Use*, 203–230. Cambridge University Press. https://doi.org/10.1017/CBO9780511804489.009.

Binder, Frank, Bastian Entrup, Ines Schiller & Henning Lobin. (2014). Uncertain about uncertainty: Different ways of processing fuzziness in digital humanities data. In *Proceedings of Digital Humanities 2014, Lausanne, 07.-11-07.2014*, 95–98. https://bit.ly/3vdP1cH.

Blaxter, Tam T. (2014). Applying keyword analysis to gendered language in the íslendingasögur. *Nordic Journal of Linguistics* 37(2). 169–198. https://doi.org/10.1017/S0332586514000171.

Bolukbasi, Tolga, Kai-Wei Chang, James Y. Zou, Venkatesh Saligrama & Adam T. Kalai. (2016). Man is to computer programmer as woman is to homemaker? Debiasing word embeddings. *Advances in Neural Information Processing Systems* 29. 1–25. https://doi.org/10.48550/arXiv.1607.06520.

Burnard, Lou. (2014). *What is the Text Encoding Initiative?: How to Add Intelligent Markup to Digital Resources* (Encyclopédie numérique). OpenEdition Press. https://books.openedition.org/oep/426?lang=fr.

Cameron, Deborah. (1995). *Verbal Hygiene*. Routledge.

Cantos, Pascual. (2014). The use of linguistic corpora for the study of linguistic variation and change: types and computational applications. In Juan Manuel Hernández-Campoy & Juan Camilo Conde-Silvestre (eds.), *The Handbook of Historical Sociolinguistics*, 99–122. John Wiley & Sons, Ltd. https://doi.org/10.1002/9781118257227.ch6.

Chambers, J. K. (1995). *Sociolinguistic Theory: Linguistic Variation and its Social Significance*. Blackwell.

Ciula, Arianna & Øyvind Eide. (2017). Modelling in digital humanities: signs in context. *Digital Scholarship in the Humanities* 32(suppl_1). i33–i46. https://doi.org/10.1093/llc/fqw045.

Connors, Richard. (1997). Poor women, the parish and the politics of poverty. In Hannah Barker & Elaine Chalus (eds.), *Gender in Eighteenth Century England: Roles, Representations and Responsibilities*, 126–147. Longman.

Craig, Hugh & R. Whipp. (2010). Old spellings, new methods: automated procedures for indeterminate linguistic data. *Literary and Linguistic Computing* 25(1). 37–52. https://doi.org/10.1093/llc/fqp033.

Crowley, Tony. (1997). Uniform, excellent, common: reflections on standards in language. *Language Sciences* 19(1). 15–21. https://doi.org/10.1016/0388-0001(95)00023-2.

Crowley, Tony. (2003). The standard language: The language of the literate. In *Standard English and the Politics of Language*, 106–137. Springer. https://doi.org/10.1057/9780230501935_5.

Culpeper, Jonathan & Merja Kytö. (2010). *Early Modern English Dialogues: Spoken Interaction as Writing*. Cambridge University Press.

Daybell, James. (2006). *Women Letter-Writers in Tudor England*. Oxford University Press.

Domingo, Miguel & Francisco Casacuberta. (2020). Modernizing historical documents: A user study. *Pattern Recognition Letters* 133. 151–157. https://doi.org/10.1016/j.patrec.2020.02.027.

Elspaß, Stephan. (2011). A twofold view 'from below': New perspectives on language histories and language historiographies. In *Germanic Language Histories 'from Below' (1700-2000)*, 3–10. De Gruyter. https://doi.org/10.1515/9783110925463.3.

Elspaß, Stephan. (2012). The use of private letters and diaries in sociolinguistic investigation. In Juan Manuel Hernández-Campoy & Juan Camilo Conde-Silvestre (eds.), *The Handbook of Historical Sociolinguistics*, 156–169. Wiley-Blackwell Malden, MA & Oxford. https://www.researchgate.net/publication/349439861_The_Use_of_Private_Letters_and_Diaries_in_Sociolinguistic_Investigation.

Epstein, Joshua M. (2008). Why model? *Journal of Artificial Societies and Social Simulation* 11(4). https://www.jasss.org/11/4/12.html.

Fairman, Tony. (2000). English pauper letters 1800-34, and the English language. In David Barton & Nigel Hall (eds.), *Letter Writing as a Social Practice*, 63–82. John Benjamins Publishing. https://doi.org/10.1075/swll.9.

Fairman, Tony. (2007). Writing and 'the standard': England, 1795–1834. *Multilingua - Journal of Cross-Cultural and Interlanguage Communication* 26(2-3). 167–201. https://doi.org/10.1515/MULTI.2007.009.

Fairman, Tony. (2011). 'Lower-order' letters, schooling and the English language, 1795 to 1834. In *Germanic Language Histories 'from Below' (1700-2000)*, 31–44. De Gruyter. https://doi.org/10.1515/9783110925463.31.

Feldman, David. (2003). Migrants, immigrants and welfare from the Old Poor Law to the welfare state. *Transactions of the Royal Historical Society* 13. 79–104. https://www.jstor.org/stable/3679247.

Fishman, Joshua A. (ed.). (1968). *Readings in the Sociology of Language*. De Gruyter Mouton. https://doi.org/10.1515/9783110805376.

Fitzmaurice, Susan. (2015). English aristocratic letters. In Anita Auer, Daniel Schreier & Richard J. Watts (eds.), *Letter Writing and Language Change*, 156–184. Cambridge University Press. https://doi.org/10.1007/9781139088275.010.

Fletcher, Anthony. (1995). *Gender, Sex and Subordination in England 1500-1800*. Yale University Press.

Görlach, Manfred. (1999). Regional and social variation (Chapter 6). In *The Cambridge History of the English Language. iii. 1476 to 1776*, 459–538. Cambridge University Press. https://doi.org/10.1017/CHOL9780521264761.007.

Gries, Stefan Th. & John Newman. (2018). Creating and using corpora. In Robert J. Podesva & Devyani Sharma (eds.), *Research Methods in Linguistics*, 257–287. Cambridge University Press. https://doi.org/10.1017/cbo9781139013734.015.

Grund, Peter. (2007). From tongue to text: The transmission of the Salem witchcraft examination records. *American Speech* 82(2). 119–150. https://doi.org/10.1215/00031283-2007-005.

Hernández-Campoy, Juan Manuel & Juan Camilo Conde-Silvestre (eds.). (2014). *The Handbook of Historical Sociolinguistics* (Blackwell Handbooks in Linguistics). John Wiley & Sons, Ltd.

Hernández-Campoy, Juan Manuel & Natalie Schilling. (2014). The application of the quantitative paradigm to historical sociolinguistics: Problems with the generalizability principle. In Juan Manuel Hernández-Campoy & Juan Camilo Conde-Silvestre (eds.), *The Handbook of Historical Sociolinguistics*, 63–79. John Wiley & Sons, Ltd.

Hill, Bridget. (2013). *Women, Work & Sexual Politics in Eighteenth-Century England*. Routledge.

Hitchcock, Tim. (2004). A new history from below. In *History Workshop Journal*, vol. 57, 294–298. https://doi.org/10.1093/hwj/57.1.294.

Kaislaniemi, Samuli, Mel Evans, Teo Juvonen & Anni Sairio. (2017). 'A graphic system which leads its own linguistic life'? Epistolary spelling in English, 1400-1800. *Exploring Future Paths for Historical Sociolinguistics. Advances in Historical Sociolinguistics (AHS)* 7. 187–213. https://doi.org/10.1075/ahs.7.08kai.

Kiełkiewicz-Janowiak, Agnieszka. (2014). Class, age, and gender-based patterns (Chapter 17). In Juan Manuel Hernández-Campoy & Juan Camilo Conde-Silvestre (eds.), *The Handbook of Historical Sociolinguistics*, 307–331. Wiley Online Library. https://doi.org/10.1002/9781118257227.ch17.

Kilgarriff, Adam. (2005). Language is never, ever, ever, random. *Corpus Linguistics and Linguistic Theory* (1-2). 263–275. https://doi.org/10.1515/cllt.2005.1.2.263.

King, Peter. (2004). Social inequality, identity and the labouring poor in eighteenth-century England. In Henry French & Jonathan Barry (eds.), *Identity and Agency in England, 1500–1800*, 60–86. Springer. https://doi.org/10.1057/9780230523104_3.

King, Steven. (2019). *Writing the Lives of the English Poor, 1750s-1830s*. McGill-Queen's University Press.

Kirschenbaum, Matthew G. (2010). What is digital humanities and what's it doing in English departments? *ADE Bulletin* (150). 55–61. https : / / www . uvic . ca / humanities/english/assets/docs/kirschenbaum.pdf.

Kytö, Merja. (2011). Corpora and historical linguistics. *Revista Brasileira de Linguística Aplicada* 11. 417–457. https://doi.org/10.1590/S1984-63982011000200007.

Kytö, Merja & Terry Walker. (2003). The linguistic study of Early Modern English speech-related texts: How "bad" can "bad" data be? *Journal of English Linguistics* 31(3). 221–248. https://doi.org/10.1177/0075424203257260.

Labov, William. (1972). *Sociolinguistic Patterns*. University of Pennsylvania Press.

Labov, William. (1994). The Use of the Present to Explain the Past. In *Principles of Linguistic Change. Volume 1: Internal Factors*, 9–27. Blackwell Publishing.

Labov, William. (2001). The Gender Paradox. In *Principles of Linguistic Change. Volume 2: Social Factors*, 261–293. Blackwell Publishing.

Labov, William. [1966] (2015). Hypercorrection by the lower middle class as a factor in linguistic change. In William Bright (ed.), *Sociolinguistics*, 84–113. De Gruyter Mouton. https://doi.org/10.1515/9783110856507.

Lahti, Leo, Eetu Mäkelä & Mikko Tolonen. (2020). Quantifying bias and uncertainty in historical data collections with probabilistic programming. In *CHR 2020 Computational Humanities Research 2020 Proceedings of the Workshop on Computational Humanities Research Amsterdam, the Netherlands, 11, 18-20, 2020.* 280–289. http://ceur-ws.org/Vol-2723/short46.pdf.

Laitinen, Mikko. (2015). Early nineteenth-century pauper letters. In Anita Auer, Daniel Schreier & Richard J. Watts (eds.), *Letter Writing and Language Change*, 185–201. Cambridge University Press.

Laitinen, Mikko & Anita Auer. (2014). Letters of Artisans and the Labouring Poor (England, c. 1750–1835): Approaching linguistic diversity in Late Modern English. In Simone E. Pfenninger, Olga Timofeeva, Anne-Christine Gardner, Alpo Honkapohja, Marianne Hundt & Daniel Schreier (eds.), *Contact, Variation, and Change in the History of English* (Studies in Language Companion Series 159), 187–211. John Benjamins Publishing Company. https://benjamins.com/catalog/slcs.159.10lai.

Lakoff, Robin. (1973). Language and woman's place. *Language in Society* 2(1). 45–79. https://web.stanford.edu/class/linguist156/Lakoff_1973.pdf.

Lawson, John & Harold Silver. [1973] (2013). *A Social History of Education in England*. Routledge. https://doi.org/10.4324/9781315887951.

Lees, Lynn Hollen. (1998). *The Solidarities of Strangers: The English Poor Laws and the People, 1700-1948*. Cambridge University Press.

López-Couso, María José. (2016). Corpora and online resources in English historical linguistics. In *The Cambridge Handbook of English Historical Linguistics*, 127–145. Cambridge University Press. https://doi.org/10.1017/CBO9781139600231.009.

Mäkelä, Eetu, Krista Lagus, Leo Lahti, Tanja Säily, Mikko Tolonen, Mika Hämäläinen, Samuli Kaislaniemi & Terttu Nevalainen. (2020). Wrangling with non-standard data. In *Proceedings of the Digital Humanities in the Nordic Countries 5th Conference*

Riga, Latvia, October 21-23, 2020. https://www.researchgate.net/publication/342110105_Wrangling_with_Non-Standard_Data.

Marquilhas, Rita & Iris Hendrickx. (2014). Manuscripts and machines: The automatic replacement of spelling variants in a Portuguese historical corpus. *International Journal of Humanities and Arts Computing* 8(1). 65–80. https://doi.org/10.3366/ijhac.2014.0120.

McCarty, Willard. (2002). Humanities computing: Essential problems, experimental practice. *Literary and Linguistic Computing* 17(1). 103–125. https://doi.org/10.1093/llc/17.1.103.

McColl Millar, Robert. (2012). *English Historical Sociolinguistics*. Edinburgh University Press.

McEnery, Tony & Andrew Hardie. (2012). *Corpus Linguistics: method, Theory and Practice*. Cambridge University Press.

Menegatti, Michela & Monica Rubini. (2017). Gender bias and sexism in language. In *Oxford Research Encyclopedia of Communication*. https://oxfordre.com/communication/view/10.1093/acrefore/9780190228613.001.0001/acrefore-9780190228613-e-470.

Meunier, Jean-Guy. (2017). Humanités numériques et modélisation scientifique. *Questions de Communication* (31). 19–48. https://www.cairn.info/revue-questions-de-communication-2017-1-page-19.htm.

Milroy, James. (1992). *Linguistic Variation and Change. On the Historical Sociolinguistics of English*. Blackwell Publishers.

Milroy, James & Lesley Milroy. (2012). Prescription and Standardisation (Chapter 1). In *Authority in Language: Investigating Standard English*, 1–23. Routledge.

Nevalainen, Terttu. (1999). Making the best use of 'bad' data: Evidence for sociolinguistic variation in Early Modern English. *Neuphilologische Mitteilungen*. 499–533. https://www.jstor.org/stable/pdf/43346227.pdf.

Nevalainen, Terttu. (2000). Gender differences in the evolution of standard English: evidence from the Corpus of Early English Correspondence. *Journal of English Linguistics* 28(1). 38–59. https://doi.org/10.1177/00754240022004866.

Nevalainen, Terttu. (2002). Language and woman's place in Earlier English. *Journal of English Linguistics* 30(2). 181–199. https://doi.org/10.1177/007242030002006.

Nevalainen, Terttu. (2018). Approaching change in 18th-century English (Chapter 1). In Terttu Nevalainen, Minna Minna Palander-Collin & Tanja Säily (eds.), *Patterns of Change in the 18th Century. a Sociolinguistic Approach*, 3–12. John Benjamins Publishing Company. https://doi.org/10.1075/ahs.8.

Nevalainen, Terttu & Helena Raumolin-Brunberg. (2017). *Historical Sociolinguistics. Language Change in Tudor and Stuart England*. Routledge.

OED online. (2021). https://www.oed.com/.

Palander-Collin, Minna, Minna Nevala & Anni Sairio. (2013). Language and identity in letters. *Ex Philologia Lux: Essays in Honour of Leena Kahlas-Tarkka*. 291–311. https://www.academia.edu/12219426/Language_and_identity_in_letters.

Partington, Alan. (2006). Aims, tools and practices of Corpus Linguistics. *Intune Papers, No. ME-06-05.* 12. https://www.academia.edu/1117163/Aims_tools_and_practices_of_Corpus_Linguistics.

Pavé, Alain. (2005). La modélisation et la simulation des objets et processus complexes. Questions scientifiques, méthodologiques et éthiques. *Natures Sciences Sociétés* 13(2). 169–171. https://www.cairn.info/revue-natures-sciences-societes-2005-2-page-169.htm.

Pettersson, Eva, Beáta Megyesi & Jörg Tiedemann. (2013). An SMT approach to automatic annotation of historical text. In *Proceedings of the Workshop on Computational Historical Linguistics at NODALIDA 2013; May 22-24; 2013; Oslo; Norway. NEALT Proceedings Series 18/ Linköping Electronic Conference Proceedings 87*, 54–69. Linköping University Electronic Press. https://www.researchgate.net/publication/257921594_An_SMT_Approach_to_Automatic_Annotation_of_Historical_Text.

Pfeffer, Avi. (2016). *Practical Probabilistic Programming*. Simon & Schuster. https://livebook.manning.com/book/practical-probabilistic-programming/about-this-book/.

Piotrowski, Michael. (2012). *Natural Language Processing for Historical Texts* (Synthesis Lectures on Human Language Technologies 17). Morgan & Claypool.

Piotrowski, Michael. (2019a). Accepting and modeling uncertainty. *Zeitschrift für digitale Geisteswissenschaften*. https://doi.org/10.17175/sb004_006a.

Piotrowski, Michael. (2019b). Historical models and serial sources. *Journal of European Periodical Studies* 4(1). 8–18. https://doi.org/10.21825/jeps.v4i1.10226.

Piper, Andrew. (2017). Think small: on literary modeling. *PMLA/ Publications of the Modern Language Association of America* 132(3). 651–658. https://doi.org/10.1632/pmla.2017.132.3.651.

Prates, M. O. R., P. H. Avelar & L. C. Lamb. (2018). Assessing gender bias in machine translation: A case study with Google Translate. *Neural Computing and Applications* 32. 6363–6381. https://doi.org/10.1007/s00521-019-04144-6.

Purvis, June. (1989). *Hard Lessons. the Lives and Education of Working-Class Women in Nineteenth-Century England*. University of Minnesota Press.

Rijswijk, Kelly, W. Bulten, L. Klerkx, L. den Dulk, Joost Dessein & Lies Debruyne. (2020). *Digital Transformation: Ongoing digitisation and digitalisation processes*. https://desira2020.eu/wp-content/uploads/2020/05/DESIRA_article_rijswijk_digitisation.pdf.

Rissanen, Matti. (2000). The world of English historical corpora. from Cædmon to the computer age. *Journal of English Linguistics* 28(1). 7–20. https://doi.org/10.1177/00754240022004848.

Rissanen, Matti. (2018). Three problems connected with the use of diachronic corpora. *ICAME Journal (Reprinted from ICAME Journal, vol. 13 (1989): 16-19)* 42(1). 9–12. https://doi.org/10.1515/icame-2018-0002.

Romaine, Suzanne. (1982). *Socio-Historical Linguistics: Its Status and Methodology* (Cambridge Studies in Linguistics). Cambridge University Press.

Salmon, Vivian. (1988). English punctuation theory 1500-1800. *Anglia - Zeitschrift für englische Philologie* (106). 285–314. https://doi.org/10.1515/angl.1988.1988.106.285.

Sauret, Nicolas. (2017). Epistémologie du modèle. Des Humanités syntaxiques? *Sens Public*. https://papyrus.bib.umontreal.ca/xmlui/handle/1866/19746.

Shave, Samantha A. (2017). *Pauper Policies. Poor Law Practices in England, 1780-1850*. Manchester University Press.

Simonton, Deborah. (2000). Schooling the poor : Gender and class in eighteenth-century England. *Journal for Eighteenth-Century Studies* 23(2). 183–202. https://doi.org/10.1111/j.1754-0208.2000.tb00586.x.

Smaldino, Paul E. (2020). How to translate a verbal theory into a formal model. *Social Psychology*. https://doi.org/10.1027/1864-9335/a000425.

Smitterberg, Erik. (2012). Late Modern English: Sociolinguistics (Chapter 59). In Alexander Bergs & Laurel J. Brinton (eds.), *English Historical Linguistics: An International Handbook*, vol. 34, 952–965. Mouton de Gruyter. https://doi.org/10.1515/9783110251593.952.

Sokoll, Thomas. (2000). Negotiating a living: Essex pauper letters from London, 1800-1834. *International Review of Social History* 45(S8). 19–46. https://www.jstor.org/stable/44735355.

Sokoll, Thomas. (2006). Writing for relief: Rhetoric in English pauper letters, 1800-1834. In *Being Poor in Modern Europe. Historical Perspectives 1800-1940*, 91–111. Peter Lang AG. https://www.fernuni-hagen.de/geschichte/lg1/docs/publikationen/sokoll_writing_for_relief.pdf.

Sun, Tony, Andrew Gaut, Shirlyn Tang, Yuxin Huang, Mai ElSherief, Jieyu Zhao, Diba Mirza, Elizabeth Belding, Kai-Wei Chang & William Yang Wang. (2019). Mitigating gender bias in natural language processing: Literature review. *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. 1630–1640. https://doi.org/10.18653/v1/P19-1159.

Tarte, Ségolène M. (2011). Digitizing the act of papyrological interpretation: Negotiating spurious exactitude and genuine uncertainty. *Literary and Linguistic Computing* 26(3). 349–358. https://doi.org/10.1093/llc/fqr015.

Trudgill, Peter. (1999). Standard English: what it isn't. In Tony Bex & Richard J. Watts (eds.), *Standard English. The Widening Debate*, 117–128. Routledge. https://vceresources.stpats.vic.edu.au/uploads/8/4/1/5/8415601/trudgill.1999.pdf.

Trudgill, Peter. (2020). Sociolinguistic typology and the uniformitarian hypothesis. In Emily Irene Crevels & Pieter Muysken (eds.), *Language Dispersal, Diversification, and Contact*, 44–57. Oxford University Press, USA. https://doi.org/10.1093/oso/9780198723813.003.0003.

Tumbe, Chinmay. (2019). Corpus linguistics, newspaper archives and historical research methods. *Journal of Management History*. https://doi.org/10.1108/JMH-01-2018-0009.

Van Zundert, Joris J. (2016). Screwmeneutics and hermenumericals. The computationality of hermeneutics. In Susan Schreibman, Ray Siemens & John Unsworth (eds.), *A New Companion to Digital Humanities*, 331–347. Wiley-Blackwell. https : / / onlinelibrary.wiley.com/doi/10.1002/9781118680605.ch23.

Vandenbussche, Wim & Stephan Elspaß. (2007). Introduction: Lower class language use in the 19th century. *Multilingua - Journal of Cross-Cultural and Interlanguage Communication* 26(2-3). 167–201. https://doi.org/10.1515/MULTI.2007.007.

Watts, Richard J. (2014). Language myths. In Juan Manuel Hernández-Campoy & Juan Camilo Conde-Silvestre (eds.), *The Handbook of Historical Sociolinguistics*, 585–606. John Wiley & Sons, Ltd.

Watts, Richard J. (2015). Setting the scene: Letters, standards and historical sociolinguistics. In Anita Auer, Daniel Schreier & Richard J. Watts (eds.), *Letter Writing and Language Change*, 1–13. Cambridge University Press Cambridge.

Weinreich, Uriel, William Labov & Marvin Herzog. (1968). Empirical foundations for a theory of language change. In Winfred P. Lehman & Yakov Malkiel (eds.), *Directions for Historical Linguistics. A Symposium.* 95–195. University of Texas Press. https : //bit.ly/3DQmdbw.

Willen, Diane. (1988). Women in the public sphere in Early Modern England: The case of the urban working poor. *The Sixteenth Century Journal*. 559–575. https://www.jstor.org/stable/2540987.

Windhager, Florian, Saminu Salisu, Günther Schreder & Eva Mayr. (2019). Uncertainty of what and for whom - and does anyone care? Propositions for cultural collection visualization. In *4th IEEE Workshop on Visualization for the Digital Humanities (VIS4DH). Vancouver, Canada*, 1–5. https://bit.ly/3kOXtc4.

Woolf, Daniel & Adam Fox (eds.). (2003). *The Spoken Word: Oral Culture in Britain, 1500-1850*. Manchester University Press.

Zhao, Jieyu, Tianlu Wang, Mark Yatskar, Ryan Cotterell, Vicente Ordonez & Kai-Wei Chang. (2019). Gender bias in contextualized word embeddings. *arXiv:1904.03310v1 [cs.CL] (Cornell University)*. https : / / arxiv . org / abs / 1904.03310.

Appendices

Appendix A

Archives

In order to respect any existing copyright restrictions, I have chosen not to include the texts of the pauper letters used for my subcorpus in this *mémoire*. The letters are, however, available and are to be found in the following archives in the UK:

List of letters written by pauper women

- Bedfordshire: BF/SH/2
- Cambridgeshire: CA/RO/1: 28, CA/SW/5: 10, CA/TR/2: 31
- Derbyshire: DB/BK/1: 4, DB/TI/2: 12r
- Durham: DU/BU/3, DU/BC/5
- Herefordshire: HE/BR/1: 9+
- Oxfordshire: OX/CL/1: 1

List of letters written by pauper men

- Berkshire: BK/PA/18: 28
- Derbyshire: DB/BK/4: 15+, DB/BS/6: 1
- Dorset: DO/WM/3
- Herefordshire: HE/EA/1: 3+

- Huntingdonshire: HU/BR/3: 8+
- Lancashire: LA(B)/MI/4
- Nottinghamshire: NG/MA/1
- Oxfordshire: OX/CL/3: 3, OX/CH/1: 4