Manual for the Scottish Pauper Petition (ScotPP) corpus

By Alma Strakova & Hester Groot
In collaboration with Mo Gordon & Jelena Prokic

Universiteit Leiden
1 September 2022

Table of contents

Foreword

Welcome to the ScotPP (Scottish Pauper Petitions) corpus manual! The Scottish Pauper Petitions corpus was created as a part of a research traineeship project at Leiden University. The project was led by Dr. Moragh Gordon and Dr. Jelena Prokic, and was worked on by Hester Groot and Alma Strakova. The aim of the research traineeship was to create a corpus of Scottish Pauper Petitions that could eventually be published. The corpus, in its current stage, is not accessible to the public; however, if you wish to access it, please contact Dr. Moragh Gordon at m.s.gordon@hum.leidenuniv.nl .

This is the instruction manual for users of the ScotPP corpus. The corpus' manual is divided in sections. The first section of the manual describes the content of the corpus, i.e. the materials it contains (as of August 2022). The second section describes the choices made in the transcription and annotation of the materials and data in the corpus, which may be helpful for future researchers who hope to make use of the materials for their own studies and/or research.

We hope you enjoy the use of the corpus. Any further questions can be posed via the contact form on our website, www.scotpp.lucdh.nl, or by emailing one of the co-creators (you can find our contact details on the website as well).

Introduction

Background to the corpus

This corpus was created as part of a project launched at Leiden University in 2022, a research traineeship under the name "Language of the Poor in Late Modern Scotland: From the Archive to the Internet". As the name suggests, this project was set up with a historical linguistic angle in mind; its aim is to, through building the ScotPP corpus, provide a source of linguistic data for a period and social group that has been little investigated prior. This group is the lower class of Scotland in the nineteenth century, a period nearing the end of Scottish English's Late Modern era.

This historical period has received little scholarly attention in the field of linguistics. While the development of Scottish English in the centuries prior has been much investigated, this research has focused largely on the initial stages of standardisation and anglicisation of the sixteenth through eighteenth century, and, specifically, on the language of upper-class writers from that period. A big reason for this is the simple matter of availability of data. Historical linguistics suffers from the Labov (1994)-termed *bad data problem*, where adequate primary sources are often not available, having been badly preserved or nonexistent. Thus, the historical linguistic research of Scottish English of this time period has been highly influenced by sources written by upper-class individuals.

However, recent research in the field has called for an approach that better reflects the multifaceted nature of linguistic developments, highlighting the language of social groups and individuals that have traditionally been underrepresented. Focusing on so-called *language histories from below*[1] offers an innovative and valuable perspective on language historiography and research.

The compilation of the ScotPP corpus is an effort to contribute towards this shift in scholarship towards the representation of previously unshown language forms. By making available innovative materials that represent the language of a new stratus of Scottish society, one which had previously been scarcely documented in linguistic and historical corpora, this corpus helps to alleviate the bad data problem by filling an important gap, and puts the histories of lower class people of Scotland front and centre.

Furthermore, the corpus is useful not only for linguistic scholars, but also for historians, as its materials offer a detailed and personal insight into the workings of the poor relief system in nineteenth-century Scotland, as well as the lives of the authors included in the corpus. While the

[1] Elspaß, St., Langer, N., Scharloth, J. & Vandenbussche, W. (2007). Germanic Language Histories 'from below' (1700–2000). Berlin: De Gruyter.

materials are, at the time of publication, limited to sources from the parishes of Tongue and Perth, the hope is that the corpus will be built on and expanded with further materials when those become available, thereby creating a geographically varied and extensive collection of lower-class writing and historical documentation, suitable for a variety of research purposes.

Materials

The materials of the ScotPP corpus consist of pauper petitions written in nineteenth-century Scotland, from a selection of parishes that at the time of writing consists of Dumfries, Glasgow, Perth, and Tongue. Pauper petitions are a form of letter-writing that played a vital role in the poor relief system of nineteenth-century Scotland. Scotland's poor law system, both before and after the 1845 implementation of the New Poor Law, required parishes to provide for their poor financially and with further necessary aid. In order to qualify for this aid, claimants generally wrote to their parish inspectors and/or the local Poor Board, detailing their situation and the reasons why they might be entitled to poor relief. Such reasons might include illness, disability, or old age; based on the petitions they received, the inspector and Poor Board decided whether or not the petitioner would be admitted to the poors' roll, and entitled to the financial support that came with that.

Pauper petitions were the most common form of petitioning, though not the only one. Jones and King (2016), in their historical study of petitioning and the Scottish poor relief system, characterise pauper petitions as a particularly "rigid" and "supplicatory" genre of letter-writing. This is due in part to the highly formulaic structures of which the petitions are made up, often requiring certain opening and closing phrases and dictating the entire form of the letter. Throughout this corpus, different degrees of rigidity and formulaic-ness were found in the letters from the different parishes. Indeed, many included letters do not conform strictly to the pauper petition genre at all, as it was established at the time (there existed manuals on how to correctly compose petitions; cf. Daybell 2012).

Moreover, differences exist in the degree to which petitioners were able to compose their letters themselves. Oftentimes, petitions would be written by scribes, whether professional or otherwise, in instances where petitioners themselves were illiterate or simply less proficient writers. Where a petition has been written by a scribe, it is generally recognisable by the sign-off of the letter. There, petitioners leave a so-called 'x-mark', which functions as their signature. Such scribe-written petitions occur frequently throughout the data, and formed an important part of the poor relief system of the time.

To create the corpus, letters were collected from four Scottish archives: the Highland Archive Service, the Perth & Kinross Archive, the Mitchell Library, and the Ewart Library. Collecting the materials from the latter two archives, which concern the Glasgow and the Dumfries letters respectfully, has not been finished, thus, this data is not included in this section. All data that is referred to in this section concerns data collected from the Tongue and Perth letters.

In total, the collected dataset has 13.209 tokens, covering two distinct time periods. The first source included letters from the parish of Tongue, written between 1850 and 1852. The second source included letters from the parish of Perth, written in 1821. It is of note that neither of these

sources exclusively included pauper petition letters. Both of the sources included other types of writing, mainly notes, most likely written by their respective parish inspectors. These notes were generally used for cataloguing the petition letters, as they added external information to the petitions regarding whether they had been accepted or denied by the local authority. Furthermore, the dataset from Tongue also included a number of short medical reports, written by the local doctor. These reports were usually written as assessments of medical claims made by the petitions, either affirming or denying them. Some of these reports also included assessments of domestic situations, or any other claims mentioned in the petitions or that had arisen during the doctor's visit. These reports were transcribed diplomatically and filed with the petition letters. Altogether, the letters from the Tongue dataset included 54 individual petition letters, 9 medical reports by doctor R.W. Black, as well as 61 cover notes written by the inspector of the petition letter catalogue. However, some of these notes could not be ascribed to a specific petition letter. In total, the Tongue dataset included 10.086 tokens.

Conversely, the Perth dataset is much smaller, consisting of only 11 petition letters and 13 cover notes, meaning that some of the cover notes lack their corresponding petition letters. In regard to the letters themselves, most, although not all, of the petitions were written by the same scribe – Alexander Mackenzie. Furthermore, the Perth dataset includes 2 documents that catalogue the court proceedings of Margaret Spence, which are believed to have been written by herself. These documents were also included in the corpus. In total, the Perth dataset is made up of 3.123 tokens.

Exact numbers are not yet available on the Glasgow and Dumfries petitions, as those are still in the process of being transcribed. However, some initial remarks can be made. The Glasgow dataset is slightly smaller than the Perth set, consisting of 9 petitions at present. The exact number of tokens is still unknown. Most of these letters were written in the 1890s, but one or two originate earlier, in the 70s and even 50s. As for Dumfries, this dataset looks to be larger still than the Tongue dataset; while the number of letters has not yet been established, it is clear that they were all published in the 1870s. One thing that already stands out about the Glasgow and Dumfries petitions is that they follow the formulaic structure of the pauper petition less rigidly than the Tongue and Perth letters did before them. This point will be elaborated on further down the line, when the data has been more carefully examined.

Overall, during the transcription of the letters, a diplomatic approach was followed, marking the transcriptions of the petitions for the physical features of the letters as well. This included marking superscript and subscript items, margin-located notes, smudged, illegible, or overwritten handwriting, and more.

Coding decisions

Orthography:
The letters were transcribed diplomatically, maintaining original spellings throughout. This included the different variants of s that occurred throughout the texts, with s/long s (ſ) transcribed faithfully.

Capitalisation:
Capitals were used wherever they occurred in the letters. While this was at times difficult to determine due to ambiguity between the capital and lowercase letter, generally it was possible to distinguish between them and transcribe capitals as such.

Subscript and superscript:
Subscript and superscript were transcribed faithfully as they occurred throughout the source texts, by making use of the following tags:

Subscript: <hi rend=”subscript”></hi>
Superscript: <hi rend=”superscript”></hi>

A distinction was made between superscript/subscript elements, and elements that were added into the text above or below the line, which were tagged as additions (see page xx).

Underline:
Underlined text was marked as such by making use of the following tag:

Underline: <hi rend=”underline”></hi>

A distinction did have to be made between lines used to underline parts of the text, and lines used as a text divider; this was ambiguous in a small number of cases, but generally easily distinguishable.

Italics:
Italicised text was marked as such by making use of the following tag:

Italics: <hi rend=”italic”></hi>

Bold:
Bold text was marked as such by making use of the following tag:

Bold:<hi rend="bold"></hi>

Punctuation:
A number of punctuation marks were coded with tags of their own in order to diplomatically preserve the punctuation of the original documents as much as possible. Ambiguity often occurred between different punctuation marks too, such as between periods and dashes in texts. The tags used for punctuation were the following:

Ampersand (&): &
Apostrophe (‘): '
Quotation mark (“): "

Abbreviation:
Wherever abbreviated forms occurred in the text, they were marked with the abbreviation tag. In most cases, common abbreviations were not expanded. The choice to mark a unit of text with the abbreviation tag was made when the unit was considered a common abbreviation that was not ambiguous in its meaning. Notable examples that were marked as abbreviations and were not expanded included “Mr”, “Mrs”, “Dr”, and “No”.

The following tag was used for abbreviations:

<abbr>No</abbr> 15

Expansion:
Wherever text had been shortened in a more ambiguous or not universally established way, the choice was made to expand the unit of text. This was done by making as much of an educated choice as possible in deciphering the shortened form, based on context of the petition, the historical context of the letter, idiosyncratic qualities of the given petitioner, and other factors. The shortened form itself was not marked as an abbreviation, as the nesting of the tags within the program was not allowed. Only the part of the text that was expanded by the transcribers of the corpus was marked with the expansion tag.

The following tag was used for expanded forms of text:

Expansion: Oct<expan>obe</expan>r

Unclear:
Wherever a character(s) or a word was not clearly legible and an uncertain decision had to be made about a transcribed symbol, this symbol would then be encoded with the “unclear” tag. In practice, this looked as follows:

<unclear reason="hand">tomorrow</unclear>

Furthermore, this tag would be specified for the reason as to why the character is illegible. These included:
“Rubbing” - illegibility due to any kind of rubbing or smudging of ink
“Hand” - illegibility due to handwriting
“Overwriting” - illegibility due to overwriting
“Damage” - illegibility due to physical damage (such as fire, tear, etc)
“Faded” - illegibility due to faded ink

Asterisk:
Wherever a character(s) or a word was illegible, it was transcribed with an asterisk for every illegible character. The asterisk(s) would then be tagged with the “unclear” tagset as an unclear symbol.

ba<unclear reason="overwriting">*</unclear>k

Deletion:
Wherever a character or a word has been striked through (or “deleted”) in the original petition, this would be transcribed with the tags:

~~if~~

Formulaic language:
Wherever a petition contained formulaic language, i.e., phrases that often appeared in multiple petitions in similar contexts, those phrases were marked as being formulaic language. For this, the following tag was used:

<seg function="formulaic" xml:id="seg_kzl_lxp_g5b">*the Honourable Members of the Parochiell Board of the parish*</seg>

Naming decisions for the TEI Header

The automatic header that was provided for transcribing the letters by the TEI guidelines included multiple subheaders; however, only a few of these have been utilised in the creation of the ScotPP corpus thus far. The following section outlines the use of the subheaders within the ScotPP corpus.

Title statement:
The title statement in the "title" section includes the name of the petitioner, if provided on the letter, as well as the date that the given petition was written in the format year-month-day. If the letter was written by somebody in regard to a specific petitioner, this would also be noted in the title statement.

Publication statement:
The "authority" section is the same for every petition, specifying SPP (Scottish Pauper Petition corpus) as the authority of the text.

The "Internal ID" section contains the internal ID created for each transcribed petition. The ID contains the following information: the name of the parish from which the certain text comes from; whether the given text is a note, a petition, or a report (specified below); the original digits from the titles of the picture(s) that were used to transcribe the given petition (specified below). Each of these units of information, as well as each original digit number of every picture used for the transcription was separated by an underscore. Thus, a general ID looks like this:

Parish_note/petition/report_pictureid1_pictureid2

Notable information for this section:
Differentiation between a note, a petition, and a report.

"note": the cover letter of a petition which, in most cases, precedes the content of the petition letter, and, on the physical petition letter, is written on the outside of the letter. The "note", or the cover letter, usually includes general information about the petition, such as the name of the petitioner, the date of the petition, and whether or not the petition has been accepted or refused (however, there are multiple cover letters that include more information and might exclude some of the previously mentioned information). An ID of a petition was denoted as a "note" only in cases where a petition did not follow a cover letter; i.e., only a cover letter was found.

"petition": an internal ID was denoted as a petition if it included a letter that was written for the local authority by a pauper or somebody writing on behalf of the pauper with the aim to be included in the Poor's roll.

“report”: an internal ID was denoted as a report if it included a letter that was written by an authority figure, usually a doctor, as a report on a legal issue concerning a specific pauper’s condition, usually including an expert opinion on the situation.

The use of digital images.
In creating the internal IDs for the ScotPP corpus, the name of the pictures of the transcribed letter was used. More specifically, original strings of digits from the titles of the images were used in order to create a unique ID. In cases where the titles of the images included the date and time that the photo was taken, only the string of digits for the time was used. For example, if a photograph of a letter was titled “IMG_20220622_102455”, only the digits “102455” (or the last 6 digits of the title of the photograph which indicate the time when the photo was taken) were used in the ID name. Thus, a given petition transcribed from this picture which comes from the Glasgow archive was given the following internal ID:

Glasgow_petition_102455

If a photograph of a letter was not based on the time and date of the photograph, but rather was a random string of numbers, the internal ID only included the unique numbers of the photographs.

Notes Statement:
The “Note” section includes any possible comments that we as the transcribers and compilers of the ScotPP corpus might have had in reference to the given letter.

Source Description:
Was not utilised in the ScotPP corpus.

Profile Description:
Was not utilised in the ScotPP corpus.

Revision Description:
Was not utilised in the ScotPP corpus.

Non-linguistic encoded information

The ScotPP corpus has three non-linguistic units of information encoded in each letter (that they appear in). These are place names, people's names, and dates. Wherever any information of this nature appeared in the text, it was marked by specific tags.

Place names:
Wherever a place name appeared in a text, it was marked by the tagset <settlement>Tongue</settlement>.
This included names of parishes, cities, street names, etc.

Person names:
Wherever a name of a person appeared in a text, it was marked by the tagset <persName>John McKay</persName>.
Full names were encoded in a single tagset, as were singular first names, last names, or names for which the first name(s) were abbreviated but the last name was not. Titles, such as "Mrs" or "Dr" were not encoded within this tagset.

Dates:
Dates were marked by the tagset <date when="1852-02-03">3rd February. 1852</date>.
Within this tagset, the element "when" was given the value of the date of the text in the format year-month-day.