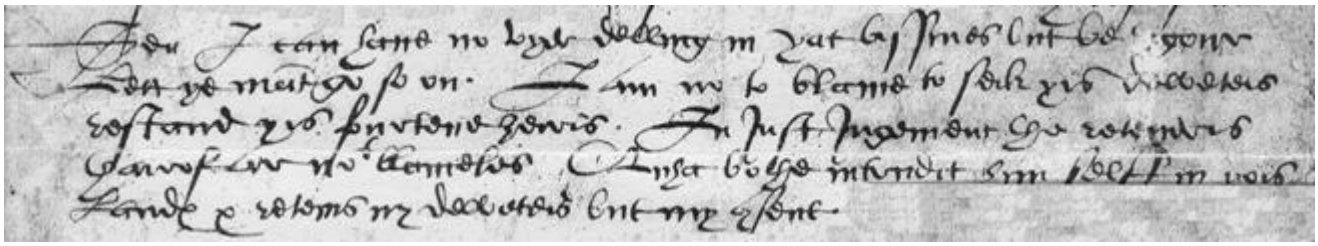
Manual to the Helsinki Corpus of Scottish Correspondence 1540–1750 (ScotsCorr)



Anneli Meurman-Solin

University of Helsinki

Helsinki 2016

An extract from a letter by William Douglas, Marquis of Angus, 1642. National Records of Scotland GD205/1/34. Published by the kind permission of the Trustees of Sir David Ogilvy of Inverquharity, Bt. (National Records of Scotland, GD205).

Abstract

Description and Specifications:

The Helsinki Corpus of Scottish Correspondence comprises circa 417,000 words of early Scottish correspondence by male and female writers dating from the period 1540-1750. Unlike the majority of digital resources available for historical linguistics at present, the corpus consists of transcripts of original letter manuscripts, which reproduce the text disallowing any modernisation, normalisation, or emendation. Language-external variables such as date, region, gender, addressee, hand, and script type have been coded into the database. The writers originate from fifteen different regions of Scotland: Aberdeenshire, Angus, Argyllshire, Ayrshire, Border counties, Fife, Invernessshire, Lanarkshire, Lothian, Moray, Perthshire, Ross, South-West, Stirlingshire, and Sutherland; these can be grouped to represent the areas of North (25 per cent), North-East (14 per cent), Central (12.5 per cent), South East (35.5 per cent), and South-West (13 per cent). There are altogether 466 informants, of which 43 remain unlocalized (c. 15,000 words). There is one category which has not been defined by the geographical origin of the writer: the parameter value Professional has been given to lawyers and members of the army and the clergy (c. 28,000 words). In addition, there is a small sample of letters by Queen Mary, James VI, and three Regents of Scotland. The proportion of female informants in the corpus is approximately 21 per cent. The proportion of relatively inexperienced and untrained writers is higher in a corpus of correspondence than in data representing most of the other genres in digital corpora available for the study of historical linguistics. Most importantly, data transcribed from manuscript without introducing modern devices for marking syntactic structure such as capitalization and punctuation ensures the validity of evidence for the reconstruction of historical syntax and discourse.

Reference

Anneli Meurman-Solin, Research Unit for the Study of Variation, Contacts and Change in English (VARIENG), Department of Modern Languages, University of Helsinki: The Helsinki Corpus of Scottish Correspondence 1540–1750 (2017) [text corpus]. - FIN-CLARIN [referred to on dd.mm.yyyy]. Available in Kielipankki, the Language Bank of Finland, at http://urn.fi/urn:nbn:fi:lb-201411071

Copyright holder of Resource

Dr Anneli Meurman-Solin, Research Unit for the Study of Variation, Contacts and Change in English. Department of Modern Languages, University of Helsinki, Finland.

It should be noted that the compiler's copyright only applies to the annotated transcripts of the letter manuscripts. The copyright of the source manuscripts lies with their repository or, in some cases, with private owners. The original manuscripts used to compile ScotsCorr are deposited in the National Records of Scotland and the National Library of Scotland, Edinburgh, UK, and, as regards the right to reproduce or to publish the manuscript sources of the historical documents themselves, the user should contact these institutions for more information.

Preface and acknowledgements

This manual describes the principles and practices applied to the selection of original manuscripts of early Scottish correspondence and their transcription and digitization for the Helsinki Corpus of Scottish Correspondence 1540–1750 (ScotsCorr). The general aim of the corpus project is to make available authentic historical data which has been reproduced in digital format by rigorously applying philologically valid guidelines as depicted, for example, by Lass (2004). While the manual hopefully provides practical information aimed at making the compiler's decisions in the transcription of the manuscripts as transparent as possible for new users of the database, auxiliary information about a number of language-external variables related to the texts and their authors and addressees will permit the use of the corpus for research especially in the fields of historical dialectology, historical sociolinguistics, and historical pragmatics (see Section 5).

In addition to the manual, there are seven auxiliary databases (see section Documentation in KORP: ScotsCorr). Three of them contain information about language-external variables defined for each letter: Male Informants in the Helsinki Corpus of Scottish Correspondence, Female Informants in the Helsinki Corpus of Scottish Correspondence, and Royal Informants in the Helsinki Corpus of Scottish Correspondence. The most convenient way of consulting these databases is to start by examining the database Word Counts by Individual and Locality, in which the letters are ordered by the variables of time and space; the filenames easily available in this database can then be used for searching the three more detailed databases. The best source for statistical data is the auxiliary file ScotsCorr Quantitative Data, in which a series of tables focusing on various aspects of the ScotsCorr permit the user to assess the representativeness of the corpus as regards time, space, and gender. Larger Regions in ScotsCorr lists the districts and localities represented in the five larger regions, North, North-East, Central, South-East, and South-West. A practical guide describing the symbols and comments used in transcribing the letters is provided by the auxiliary database Symbols and Comments in ScotsCorr.

The idea behind the compilation of the *Helsinki Corpus of Scottish Correspondence* (ScotsCorr) draws on a long-term exchange of ideas between researchers active in the scholarly community of the International Computer Archive of Modern and Medieval English (ICAME) (http://icame.uib.no/) and on team-work in the Research Unit for the Study of Variation, Contacts and Change in English (http://www.eng.helsinki.fi/varieng), funded by the Academy of Finland and the University of Helsinki, in the area of compiling diachronic corpora. The following corpora could be mentioned here: the *Helsinki Corpus of English Texts* (http://www.helsinki.fi/varieng/CoRD/corpora/HelsinkiCorpus/index.html and Rissanen & Tyrkkö 2013 http://www.helsinki.fi/varieng/journal/volumes/14/rissanen_tyrkko/, the *Helsinki Corpus of Older Scots* (http://www.helsinki.fi/varieng/CoRD/corpora/HCOS/index.html), the *Corpus of Early English Correspondence Sampler* (http://www.helsinki.fi/varieng/CoRD/corpora/CEEC/ceecs.html) and Nevala & Nurmi 2013 http://www.helsinki.fi/varieng/journal/volumes/14/nevala_nurmi/, the *Corpus of Early English Medical Writing* (http://www.helsinki.fi/varieng/CoRD/corpora/CEEM/index.html and Taavitsainen & Pahta 2013 http://www.helsinki.fi/varieng/journal/volumes/14/taavitsainen_pahta/), and the *Parsed Corpus of Early English Correspondence* (PCEEC), (http://www.helsinki.fi/varieng/CoRD/corpora/CEEC/pceec.html). However, while the above-named corpora are primarily based on editions, the ScotsCorr consists exclusively of diplomatically transcribed and digitized versions of original manuscripts of letters, and therefore closely resembles

the corpora which function as databases for the Edinburgh historical atlases, the *Linguistic Atlas of Early Medieval English* (LAEME, compiled by Margaret Laing; http://www.helsinki.fi/varieng/CoRD/corpora/LAEME/index.html), covering the period c. 1150 to c. 1300, and the *Linguistic Atlas of Older Scots* (LAOS, compiled by Keith Williamson), phase 1, c. 1380 to c. 1500 (http://www.lel.ed.ac.uk/ihd/laos1/laos1.html).

The creation of coherence in the theoretical and methodological approaches of the LAEME, LAOS and ScotsCorr databases has required close long-term collaboration. LAEME and LAOS are concerned with the reconstruction of the diatopic-diachronic patterns of the medieval Anglic vernaculars of England and Scotland. The basic methodology applied to these atlases derives from that used to make A Linguistic Atlas of Late Medieval English (LALME, McIntosh, Samuels & Benskin 1986; see also http://www.helsinki.fi/varieng/CoRD/corpora/eLALME/index.html). However, the methodology created for LALME has been developed further, so that the databases of linguistic material are lexico-grammatically tagged corpora of full texts, diplomatically edited, rather than questionnaire-delimited sets of isolated word-forms (Williamson 1992/93; about historical dialectology, see Williamson 2012). Furthermore, the "fit-technique", a method of interpolating texts of unknown provenance into a dialect continuum, has been computerized (Williamson 2000, Laing &Williamson 2004).

The compilation of a corpus of Scottish correspondence was motivated by my awareness that royal, official, and family letters were a data source with unique properties for research that seeks to reconstruct both past language use and social and cultural practices (Section 2.1 Reconstruction of text languages). Correspondence can be considered a unique source in the sense that it offers both linguists and historians a wide range of informants representing different degrees of linguistic, stylistic, and socio-cultural literacy; the idiolects and group-lects also reflect the influence of geographical and social distance and mobility (Section 2.3 Letters as a data source).

A number of other factors influenced the decision-making process during the creation of the ScotsCorr (see Section 2.2 Digital data sources for Scots). Since three geographical areas are well represented in the *Corpus of Early English Correspondence* (CEEC), East Anglia, London, and the North of England, the focus on Scotland seemed very relevant. (In the CEEC corpora, the Court has been defined as a fourth area, more social than geographical; for more information on the CEEC corpora, see Nevalainen & Raumolin-Brunberg 1996 and Nevalainen & Raumolin-Brunberg 2003.) In order to trace the diachronic developments and diffusion of numerous linguistic features in the history of English, directly comparable data originating from the various areas of Scotland is required. Since the ScotsCorr comprises approximately 417,000 words of running text representing royal, official, and family letters (see Section 4.1 Selection of data for the ScotsCorr), it is comparable to the *Corpus of Early English Correspondence Sampler* (CEECS) (http://www.helsinki.fi/varieng/CoRD/corpora/CEEC/ceecs.html). However, since language use in early Scottish letters is strongly conditioned by the writers' geographical and social mobility and the types of social network they are involved in, rather than just their geographical origin, the corpus data have not been translated into a linguistic atlas (Meurman-Solin 2000a-c, 2001a). Thus, information has been provided about the geographical area the writers originate from and the place where a particular letter was written, but in order to define the variables of social mobility and socio-economic distance, the user will have to consult research on Scottish political, social and economic history, including research on literacy (e.g., Houston 1985).

I would like to thank Doctor Margaret Laing (University of Edinburgh), Professor Roger Lass (University of Capetown), and Doctor Keith Williamson (University of Edinburgh) at the Institute for Historical Dialectology (now Angus McIntosh Centre for Historical Linguistics), University of

Edinburgh, for permitting me to benefit from their unique expertise in the creation of manuscript-based diachronic corpora. A contract between the universities of Edinburgh and Helsinki allowed me to use software created by Doctor Keith Williamson in the tagging of the CSC 2007 texts. I would like to acknowledge his very important role in the process of developing the theoretical and methodological approach applied to the CSC 2003 and 2007. I would also like to thank members of the Research Unit for the Study of Variation, Contacts and Change in English (VARIENG) for their unfailing support. Without the research assistants provided by the Varieng Research Unit it would not have been possible to complete this project; Johanna Lahti, Ulla Paatola, Elina Sorva († 2006), Riikka Tuomi, Turo Vartiainen, and Minna Åkerman participated in transcribing the manuscripts. Elina Sorva was responsible for a major part of the transcription and digitization work, and also achieved a high level of expertise as a tagger during the more than three years that she participated in the project. In addition to tagging, Turo Vartiainen assisted me in the writing of the CSC 2007 manual, and Olga Timofeeva helped in tagging and the final editorial work. I am greatly indebted to the Helsinki Collegium for Advanced Studies (University of Helsinki) for funding my research during the period 2002-2007. Saara Paatero-Burtsov, a research assistant at the Collegium, transcribed a considerable number of manuscript letters in 2002-2003, and Jenni Laitinen and Eeva Hohti helped me in the creation of the auxiliary databases. Tuuli Tahko converted the CSC 2007 manual to html.

The transcription of the letters is mostly based on xerox-copies of the original documents sold by the National Records of Scotland and the National Library of Scotland, Edinburgh. This was because the compiler was able to pay only a restricted number of visits to Edinburgh. Frequently, the quality of these copies turned out to be less than satisfactory, usually because of the poor quality of the originals, which were often torn or damaged by damp. When the data were collected, the visitors themselves were not allowed to use their own cameras for taking digital images. In numerous cases, the compiler was able to recheck the manuscripts, but the reordering of the same manuscript was not always possible in the archives. Lack of funds also prevented the compiler from ordering a copy of the cover page where the address was, so that this text may be missing in some transcripts.

I am greatly indebted to the most welcoming, generous, and well-informed staff of both the National Records of Scotland and the National Library of Scotland in Edinburgh for all their kind and invaluable assistance during this long process of compiling the ScotsCorr corpus.

1 Introduction

The general purpose of the ScotsCorr corpus is to offer the international academic community a tool for both teaching and research which will permit the study of a wide range of letters representing different types of speech, discourse and text community in sixteenth-, seventeenth-, and early eighteenth-century Scotland (see Section 3 Dimensions of space, time, and social milieu). This new tool has been designed to function as a useful data source especially for historical dialectology, historical sociolinguistics, historical pragmatics, and historical stylistics, but it will also provide a rich resource for topics such as political and socio-economic history, cultural studies, women's studies, genealogy, and, if also consulting the original manuscripts, the history of Scottish handwriting (for more information on handwriting, see Meurman-Solin 2013a-c). Since this manuscript-based corpus also presents a coherent view on how methods of philological computing

can be applied to historical documents in modern corpus linguistics, it may be used as one of the standard tools in courses on linguistic and literary computing and manuscript studies (see Section 4.2 Transcription and digitization).

The ScotsCorr is the third corpus in a series of corpora comprising Scottish correspondence, the first and the second dating from 2003 and 2007 respectively. Unlike the CSC 2007, described in http://www.helsinki.fi/varieng/csc/manual/), this extended version of the 2007 corpus has not been grammatically tagged. Instead, the ScotsCorr corpus (2017), available via CLARIN, can be searched by using Korp, a Corpus Workbench-based tool (http://urn.fi/urn:nbn:fi:lb-2016121607).

In addition to information on the compilation, digitization, and tagging principles and practices applied to the database, the *Manual for the CSC 2007 corpus* contains a full description of the theoretical approach used, which is reflected in how variation, variability and change have been conceptualized, and of what implications this has for the system of lexico-grammatical tagging applied to the data. The distinctive profile of the CSC database has been created by applying a variationist approach to the tagging of linguistic and, to some extent, non-linguistic features. Instead of tagging and parsing systems in which a restricted set of conventional category labels are used to classify linguistic items word by word, either by word-class or syntactic function, the general approach in the CSC draws on principles of notional grammar, emphasizing phenomena such as categorial fuzziness and polyfunctionality, indicating potential for membership on a particular cline – one depicting nouniness or adverbhood, for example – and signalling relations between the constituent parts of collocates. See Section 3.2 Principles of tagging and Section 3.3 Practices of tagging in the CSC manual. See also Meurman-Solin 2007 http://www.helsinki.fi/varieng/journal/volumes/01/meurman-solin/, which discusses how the fuzziness and polyfunctionality of linguistic categories can be dealt with in tagging data that reflect variation and change over time. A system of this kind is particularly relevant in tagging certain language varieties, such as the idiolects of less-trained and inexperienced female writers in early Scotland, in which the influence of standardizing trends is barely visible.

Section 6 Visual prosody examines the digitization of features of visual prosody present in the manuscript originals (i.e., non-linguistic features such as manuscript layout, paragraph structure, punctuation, particular character shapes, and spacing). The typology of the transcriber's comments on features of visual prosody is illustrated in section 6, but a detailed description is available in the auxiliary databank Symbols and Comments in the ScotsCorr. Digital photographs of a number of manuscripts have been used as illustrations in Meurman-Solin 2013a and b, but the present version of the ScotsCorr does not contain digital images of the manuscripts.

1.1 General principles of corpus compilation

1.1.1 *Protean corpora: multidimensionality, flexibility, and transparency*

OED
Protean, A.
adj. **a**. Of or pertaining to Proteus; like that of Proteus; hence, taking or existing in various shapes, variable in form; characterized by variability or variation; variously manifested or expressed; changing, varying.
proteanism *n.*,
capacity for change; changeableness, variability.

In addition to the research environment at the Research Unit for the Study of Variation, Contacts 99999and Change (VARIENG), University of Helsinki, the *Helsinki Corpus of Scottish Correspondence* was created in close cooperation with the Institute for Historical Dialectology (IHD) (now Angus McIntosh Centre for Historical Linguistics) at the University of Edinburgh, where there is long-standing expertise in the creation of linguistic atlases of text languages (http://www.ppls.ed.ac.uk/lel/groups/institute-for-historical-dialectology). The tagged databases produced as part of the Edinburgh-Helsinki collaboration were planned to represent a new genre of electronic corpora. Firstly, they would be Protean in the sense that they could be continually revised and expanded. Secondly, building on the basic format of a particular corpus, the data could be manipulated into a virtually unrestricted number of structures, "research shapes", in order to achieve the best possible validity and relevance for specific, user-defined investigations. In other words, a corpus re-shaped for a particular study might contain only those parts of the base corpus which the user considered appropriate for dealing with a particular research question. The user of the ScotsCorr corpus is also invited to assess carefully the relevance and validity of the various historical documents in the study of a particular research topic.

In fact, unevenness as regards the validity and relevance of data is an inherent quality of most electronic databases, even though this may not be explicitly indicated or carefully explained in the manuals. For example, some parts of a database may be shown to be more valid than others in terms of language-external criteria, and the user should perhaps exclude texts which may weaken scholarly argumentation in a particular study (on the assessment of representativeness, see Meurman-Solin 2001a).

Thirdly, proteanism is reflected in the tagging system tailored for the CSC 2007 corpus (see Section 3.2 Principles of tagging and Section 3.3 Practices of tagging in the CSC 2007 manual). Since information in the tags is hierarchically ordered, it is possible to decide what degree of specification or refinement is required for a particular search. In the tagged corpora (see the LAEME and LAOS sites), the user can also rearrange this information or re-tag the linguistic features under investigation by using the software provided on the sites. Thus, the tags allow refinement and enrichment, so that, even though the basic information in the tag is structural and semantic, the additional information provided about contextual properties and the commentary also permit the user to search for syntactic and textual features.

A good balance between the type of research question asked and the type of data retrieved can be created by compiling transparent, flexible, and multi-dimensional corpora. Transparency in a corpus allows the user to assess carefully and critically the validity and relevance of each text with regard to specific user-defined linguistic investigations. Flexibility in a corpus allows the user to manipulate the data into a specific form in order to achieve the best possible fit between the data and the theoretical and methodological approach. Multi-dimensionality in a corpus allows the user to restructure the data by re-creating an appropriate frame of reference based on how language-external variables have been conceptualized and defined.

1.1.2 *A corpus and its various forms*

As stated above, Protean databases can be reshaped and restructured by the user to achieve the best possible validity and relevance for the study of a specific research topic. While the earlier-generation corpora can perhaps be seen as carefully structured end-products of compiler-defined corpus compilation projects, Protean corpora are databases of digitized texts which, in addition to the basic format, can also exist in various user-defined forms.

According to this approach, the basic format of the corpus remains separate from the various "research forms". However, even the basic format may undergo changes as part of the ongoing process of development, being revised and expanded at more or less regular intervals. The Protean character of a database of this kind draws on a kind of flexibility which is achieved as follows: it is possible to take a copy of all or part of the basic format and alter the tagging itself, or add to the tagging information at any level from word-morpheme upwards. Flexibility of this kind is an important asset in the study of syntax in particular, and indispensable in diachronic studies. It is also possible to add information that relates to extra-linguistic factors and analyse, define, or group these factors in a different way. Users may want to create a form which comprises strong witnesses only, with the degree of strength assessed by how precisely a language-external variable, or a set of them – features related to textual history or those related to informants, for instance – can be defined (see Section 2.1 Reconstruction of text languages and Section 3 Dimensions of space, time and social milieu).

In principle, reshaping of this kind is also possible with many of the existing corpora, but in their cases the reshaping is restricted primarily to the selection and classification of texts according to language-external variables. For example, the user may wish to research academic prose written by women in the age-range of 40 to 60 in the state of California, and he or she may then proceed by extracting a sub-corpus out of existing electronic corpora. However, proteanism is more deeply integrated in multi-dimensional corpora, as all aspects of these may be reshaped and redefined; perhaps the most important difference is that in creating a new form, new knowledge, whether related to texts, informants, the tagger's grammar, or language-external variables, can be immediately keyed in. It may be useful that the user can autonomously define what he or she considers to be relevant knowledge. Thus, compilers provide users with as much information as is available to them, but each corpus user can then critically examine the implications that information has for the specific study at hand. As I have argued elsewhere (Meurman-Solin 2001a), in my view, over-structuring corpora by using hypothetical knowledge in the definition of language-external variables is not useful. Each user should formulate the definitions in full accordance with his or her theoretical and methodological approach. In general, the creation of a separate – non-integrated – database containing information about the texts is welcome, as the availability of more detailed information about texts allows the application of a less compartmentalized and more scalar way of conceptualizing language-external variables in corpora (see Section 3 Dimensions of space, time and social milieu).

I see the compilation of Protean corpora as an ongoing process, with dynamic interaction and critical reanalysis between the stages of compiling, experimenting, revising, restructuring, and expanding. The most important requirement in creating a multi-dimensional corpus is that the principles regulating the structure, as well as those guiding compilation, digitization, and annotation practices, are as transparent as possible. Transparency is enhanced, firstly, by introducing some degree of hierarchical ordering into how language-external variables have been conceptualized. Time and space are central, since the creation of a frame of reference consisting of diachronically- and diatopically-anchored texts is essential for the identification of valid information that can be used to position other texts in the linguistic and extralinguistic worlds reconstructed in the corpus. Ideally, in addition to being localizable and having a particular date, anchor texts will be spread relatively evenly over time and space and have a relatively fixed social and communicative function; in addition, it will be possible to reconstruct the profiles of the writers, whether members of discourse communities such as professional coalitions or individuals whose language is only available in private documents, on the basis of reliable, preferably direct, evidence. In the Edinburgh Institute for Historical Dialectology (recently renamed Angus McIntosh Centre for Historical Linguistics), the notion of "primary witness" is used where the LALME refers to anchor

texts. A primary witness is a text that can be localized on the basis of *prima facie* extra-linguistic evidence of an association with a place and given date – where there is (for the time being) no contradictory evidence. That is, any anchor text or primary witness is a working hypothesis. A "secondary witness" is a text which is localized linguistically, as it lacks any or sufficient extra-linguistic indications of its provenance and/or date (Williamson 2000, 2001, Laing & Williamson 2004).

The notion of witnesses is useful, as it allows the possibility of new evidence that may alter the strength of the case for localizing a text and, indeed, the shape of the corpus – the pattern in which texts float in their multi-dimensional space cluster. The idea is that texts will be positioned in the multidimensional space of a corpus world according to a set of coordinates that have been defined either in binary terms, i.e., in terms of a dichotomy, or, if possible, through use of a scalar system. In principle, a text is thus not a permanent member of a specific group or category, filling a slot in the compiler's schema. A text is floating in the corpus space and can be fixed for the purpose of a specific study by showing that there is a valid relation between the text, the language-external variables defining it, and the research question. The user of the corpus may see the various dimensions in terms of a hierarchical system, finding some of them particularly relevant and others marginal or not valid for a specific research hypothesis. The user may also see some dimensions as more closely interrelated than others; he or she may claim that some binary variables are independent, while others form a network in which the conditioning effect of one is dependent on the converging effect of another.

Thus, I would like to suggest that the fourth generation of corpora will combine three important properties. Firstly, we define language-external variables rigorously, benefiting from information provided by various interdisciplinary forums. Secondly, we see corpora as consisting of sub-corpora that are defined not in terms of time-periods, for instance, but in reference to degrees of validity and relevance as regards their usefulness for the study of a specific research question. Thirdly, instead of marketing corpora as completed products, we see the compilation as an ongoing process, and therefore view expansion and revision as inherent characteristics of this work.

I see these three properties as interrelated. Our understanding of the complex nature of language-external variables has increased, so that we are more aware of their scalar nature, for instance, and, as a result, find some of the traditional category labels less useful, sometimes even misleading. While some variables can be defined quite precisely, others are still based on hypotheses or knowledge which draws on as yet only partially-reconstructed stages of social, cultural, and economic history. Research questions can be examined using data with varying degrees of relevance, depending on how thorough our knowledge is of specific language-external factors.

1.1.3 *Transparency of the theoretical and methodological approach*

In addition to the careful assessment of whether the basic form of a corpus provides relevant data for the study of a specific topic, an assessment of the validity of the corpus is also necessary in order to ensure that there is no theoretical and/or methodological contradiction between the approaches of the corpus compiler and the corpus user. It is perhaps not altogether unjustifiable to ask whether methods developed by modern sociolinguistics, dialectology, or discourse stylistics, for instance, can be applied to data that has not been compiled with the theoretical framework of these fields of study in mind. Perhaps the best way to illustrate what I mean is to refer to the example of a corpus that has been structured to rigorously reflect recent theoretical and methodological developments in historical sociolinguistics. I consider the *Corpus of Early English Correspondence* to be such a corpus (Nevalainen & Raumolin-Brunberg 1996, 2003).

In my own corpus-linguistic work, I ask what theoretical and methodological implications different text annotation systems might have on inventories of particular linguistic features when applied to digital databases. How is our ability to understand linguistic systems affected by the use of quasi-automatic taggers and parsers which may re-establish and re-distribute conventionalized ways of understanding and analyzing and categorizing linguistic data? Ideally, tags should guide us towards the reassessment of our criteria for linguistic categorization, rather than provide data as categorized by criteria based on preconceived properties of linguistic features (for information on flexibility and transparency in the CSC 2007 tagging system, see Section 3.2 Principles of tagging and Section 3.3 Practices of tagging in the CSC manual; see also Meurman-Solin 2007a and b).

The elaborated tagging system in the CSC 2007 aims to be agnostic with respect to schools of modern formal syntactic theory. Attempts to revise the guidelines for philological computing has been motivated by the following observations: while electronic databases have constantly improved as regards their quantitative and qualitative validity and relevance, compromises have sometimes been made in tagging by relying on pre-corpus-linguistic descriptions, resorting to automatic (i.e., non-interactive) tagging, or imposing neat category labels on the data. The main principle in the CSC 2007 tagging system is that as little linguistic theory should be integrated into a tagged corpus as possible. In other words, the tagging in the base corpus should, as far as is possible, remain neutral with respect to formal theories, particularly those of syntax, as tags reflecting assumed syntactic properties will inevitably suggest membership in a preconceived grammatical system. Meurman-Solin (2004a: 187) summarizes the principles for tagging connectives in the CSC as follows:

> this tagging system aims at indicating item-specific or collocate-specific structural features which have been interpreted as having *semantic potential* to indicate relations between clauses, irrespective of degree of grammaticalization. The rationale for not providing information about syntactic properties is that these are interpreted as secondary, while structural and semantic properties are considered primary. In other words, the core function of structural and semantic information is descriptive – descriptive at the micro-level – while that of syntactic information is interpretative – interpretative at the macro-level, i.e., intended to identify grammatical rules and constraints. The description provided by the tags may contain information on various levels of language use, including discoursal and textual features.

In cases in which it has not been possible to avoid theory-specific practices, these must be made as transparent as possible by providing detailed information about the tagging principles and practices. The main function of the tags is that they permit the creation of comprehensive inventories which are valid for the study of variation and change. Thus, they ensure reliable data searches, rather than supporting a particular grammatical analysis.

The principle of transparency is also applied to the way in which ambiguous instances have been tagged. As discussed in more detail in Section 3.2 Principles of tagging and Section 3.4 Commentary in the CSC 2007 manual, categorial fuzziness and polyfunctionality are dealt with by using a cline of co-ordinates reflecting the different readings in the grammel of a tag. Thus, instances of *any man* can be integrated into the inventory of indefinite pronouns by positioning the term on the cline of nouniness and pronounhood using the co-ordinates “n-pn”:

> $any/pn-aj>n-pn_ANY
> $man/**n-pn**<pn-aj_MAN

Ambiguity can also be indicated by a comment which makes the alternative readings explicit:

> $beseek{cause}{lat}/vpsp{indep}_*BESEIK+ING $/vpsp{indep}_+ING
> $also/av_ALS
> **{zero that&Oinf}**
> $/T_THE
> {\}
> $eternal/aj_ETERNAL
> '_GOD
> $have{n}/vsjps13<cnp+{nom}>pr-cj_HAIF $/vsjps13<cnp+{nom}>pr-cj_0
> $/P02G_zOUR
> $grace/n{ho}_GRACE
> $in/pr-cj<v_IN
> $keep/vn{rc}-av_KEIP+ING $/vn{rc}-av_+ING

The comment {zero that&Oinf} specifies the two possible readings of the complement of the verb *beseek*. The alternative consisting of a nominal *that*-clause object with *that*-deletion has been randomly chosen as the one presented first. This order is reflected in how the rest of the clause elements have been tagged, i.e., the predicate verb of the proposed *that*-clause is analysed as a present subjunctive. The other alternative is the reading of the complement as a Latinate object + bare infinitive construction. Choosing just one of these alternatives would mean imposing a particular grammatical analysis and ignoring the other. See also Section 3.4 Commentary in the CSC 2007 manual.

Attributes such as ‘Protean’, ‘multi-dimensional’, ‘flexible’ and ‘transparent’ usefully remind us of the risks of objectifying language varieties in the compilation of corpora in the way that the dictionary industry often does (Benson 2001: 21). I have discussed these risks elsewhere (Meurman-Solin 2004b), so I will just summarize some of my comments here. As also pointed out in Section 2.1 Reconstruction of text languages, there is a tendency to objectify or reify regional varieties, assuming that they form relatively homogeneous, even relatively self-contained, entities or systems; to historicize them by emphasising socio-political rather than linguistic factors and by presenting these factors as legitimizing the naming and describing of regional varieties in a certain way; and to create hierarchies, analysing a regional variety chiefly in reference to a standardized variety or adopting the comparative method in examining less prestigious varieties (Milroy 1999). See also Williamson 2004.

These tendencies may regulate processes of analysis by which linguists are trying to identify some order in heterogeneity, i.e., some relatively consistently preferred practices in data that otherwise chiefly give evidence of heterogeneity and continued variation. Attempts to demarcate areas as the territories of specific varieties may divert our attention from the examination of ordered heterogeneity which can be observed only by crossing such artificial boundaries. By defining a text community in terms of which written texts verifiably had a social and communicative function among the literate members of that community, and by using such a representative compilation of texts as data, it is possible to deobjectify and dehistoricize a language variety. To refer to Scots as an example, a particularly important consequence of de-reification in the description of this language is that variation and variety resulting from contact between varieties and languages on Scottish soil will be given due attention.

2 Scots and the reconstruction of its use in correspondence

2.1 Reconstruction of text languages

The term "Older Scots" is used to refer to language varieties in Scotland in the period from the mid-fourteenth century up to the end of the seventeenth century. The established periodization is as follows:

Pre-literary Scots	to 1375
Early Scots	1375–1450
Middle Scots	
Early Middle Scots	1450–1550
Late Middle Scots	1550–1700

Thus Older Scots is a "text language". Fleischman (2000: 34) suggests that

> [t]he term "text language" is intended to reflect the fact that the linguistic activity of such languages is amenable to scrutiny only insofar as it has been constituted in the form of extant *texts*, which we might think of as its "native speakers", even if we can't interrogate them in quite the same way as we can native speakers of living languages. Another crucial difference between text languages and living languages is that the data corpus of a text language is finite; new data only become available when previously unknown documents are discovered, whether in the form of manuscripts, printed texts, tablets, etc.

As stressed in Meurman-Solin (2004b), recent advances in corpus linguistics no longer justify the polarization of the two approaches to the reconstruction of the languages of the past, namely 'the essentially data-driven and data-oriented' approach and 'the theoretical approach', which 'extrapolates from the data in order to identify general principles and mechanisms of language change' (cf. Fleischman 2000: 34). Instead, with major advances in historical corpus linguistics, the integration of the two approaches permits us both to provide a comprehensive description of linguistic features and to identify patterns reflecting systemic developments, and ultimately to model language variation and change in a theoretically relevant way.

As with many other pre-1500 varieties of text languages, in Older Scots the scarcity of data which are representative of a sufficiently wide range of language use causes some problems. As Johnston (1997:48) points out, until recently the evidence scholars had to rely on was mostly 'documents written by a small, unrepresentative section of Older Scots society, often a specialized, "set-piece" type of text such as a will, a deed, a public record or a literary work'. With the creation of the *Edinburgh Corpus of Older Scots* database (c. 1380 to c. 1500), compiled to create the *Linguistic Atlas of Older Scots* (LAOS) (http://www.lel.ed.ac.uk/ihd/laos1/laos1.html), the situation has improved considerably, but it is not possible to create a fully balanced corpus of this variety for the pre-1500 period (Section 3 Dimensions of space, time and social milieu). There are also significant gaps in the data for the first half of the sixteenth century.

The ScotsCorr corpus aims to permit the user to detect a wide range of variation and variability in Older Scots by including idiolects of professional writers and other writers with university education, as well as less-trained and inexperienced writers, and by rejecting all practices of normalization and standardization in transcribing and digitizing the manuscript texts. Aitken (1971) sees a high degree of variation as an inherent feature of Older Scots, and his view has been amply supported by recent corpus-based research. The reconstruction of variation and change has also benefited from the fact that, in corpora of Scots, the proportion of texts written by women has increased (Meurman-Solin 2001b, 2005).

As discussed in Meurman-Solin (2004a), the reification, or objectification, of the Scottish variety and its description as part of a hierarchized system of varieties has tended to divert our attention from the more challenging task, that of providing a comprehensive description of variation and change in the various areas of Scotland. However, diatopically representative data selected from manuscript evidence makes it possible to examine the history of Scots without reference to standardization – or Anglicization (Devitt 1989) – or indeed a preconceived language system. In the present, emphatically data-driven approach, a comprehensive description can be presented more traditionally by illustrating the attested patterns of continued variation or by resorting to methods made possible by new technology.

In my view, the reconstruction of the history of Scots seems to be negatively affected by three tendencies in earlier research on such primarily geographically- and politically-defined varieties of English as Scots (Meurman-Solin 2004a). There is a tendency to objectify or reify regional varieties, assuming they form relatively homogeneous – perhaps even relatively self-contained – entities or systems; a tendency to emphasize socio-political rather than linguistic factors in order to legitimize the naming and describing of regional varieties in a certain way; and a tendency to create hierarchies, leading to the analysis of a regional variety exclusively with reference to a standardized variety (cf. Milroy 1999). The use of quantitatively- and qualitatively-improved data is necessary for the creation of an unbiased comprehensive account of language varieties which until now have been described less fully than the standardized varieties used by wider speech and text communities.

In addition to the negative implications of reification, hierarchization and historicization, the categorization of texts in a database may influence the way in which we interpret the findings. As pointed out in Meurman-Solin (2001a), language-external variables used in structuring electronic databases may lead to a compartmentalization of texts into subcategories which, when examined more closely, are internally quite complex and heterogeneous. In the ScotsCorr, no categorization into subgenres of correspondence has been provided. Instead, the user is invited to proceed from the idiolectal level to the local, regional, supraregional, and national levels, reconstructing variation and change over time and space before examining the conditioning of other language-external variables (Section 3 Dimensions of space, time and social milieu). Earlier research has shown that factors such as social status and networks play a major role (e.g., Meurman-Solin 2001b), but style- and discourse-related variables reflecting contemporary politeness strategies can also be shown to influence the choice of linguistic features in various ways (Meurman-Solin 1993, 2002, Meurman-Solin & Nurmi 2004, Meurman-Solin & Pahta 2006).

The reconstruction of past language use through historical documents encounters problems of various kinds. In addition to the above-mentioned gaps in the evidence caused by scarcity of texts, those witnesses we do have represent different degrees of validity and relevance. Depending on the type of research question, the texts in a corpus must be categorized into primary and secondary witnesses on the basis of what is known about their history, their writers, and the circumstances of

their production and distribution (cf. Nevalainen & Raumolin-Brunberg (1996: 43) and the discussion in Meurman-Solin 2001a).

As pointed out in Meurman-Solin (2004b and c), for a text to function as a primary witness or an anchor text, its history must be able to be reliably recovered, and there must also be sufficient information about its writer. Even relatively well-known anchor texts may be complex, in the sense that no straightforward claims about correlation between language-external factors and linguistic features can be made. For example, legal texts, especially public documents, usually offer good *prima facie* evidence for localization in "space" (Williamson 2000). However, despite the precise date of production, for defining the variable of "time" as a conditioning factor, it is necessary to investigate what the role of conventions and formulae is in a text, as such fixed expressions increase the general degree of conservatism in legalese. In contrast, the date of a letter allows us to specify a particular point of time in a person's idiolect in a particular communicative situation, but the definition of the variable of "space" calls for a scalar system of parameter values when applied to texts by geographically and socially mobile writers. Localization is often difficult in the case of women who, as a result of marriage (or marriages), moved from one place to another. Moreover, as we do not usually know where and when a female informant became literate, it may be impossible to tell which area the spelling practices she learned belong to and how they relate to her pronunciation.

In my earlier research (Meurman-Solin 2000a-c, 2002) I have also discussed a number of other factors which complicate the process of interpreting linguistic findings. I would like to highlight the important role of different degrees of linguistic and stylistic competence (for illustrations, see Meurman-Solin 2001b, Meurman-Solin & Nurmi 2004). Another important aspect to consider is that, although the network of family castles scattered around on the map of Scotland may give the impression of places on the periphery or in isolation, their distance from administrative centres varies depending on a particular family or family member's role in national politics, the economy, or culture.

In addition to the geographical distance between the various places of origin of the texts and their writers, I have found it useful to apply the concepts of economic and social distance in commenting on differences between members of self-contained tightly-knit speech communities and those regularly in contact with people originating from various other areas within a rather diffusely-patterned administrative and economic framework (Meurman-Solin 2000a-c). For information on the concepts of speech community, discourse community, and text community, see Section 3 Dimensions of space, time and social milieu.

To sum up the main points, the ScotsCorr database provides new information for a comprehensive descriptive account of the continuum from idiolectal and local to regional and national varieties of Scottish English in the period 1540–1750. As regards individual informants, our ability to relate our linguistic findings to language-external factors depends on how successfully we have managed to define features such as degree of geographical and social mobility, which can be estimated by drawing on information provided by social, economic, and cultural history as well as demography. In addition to basic information about literacy in Scotland (see Marshall 1983, Houston 1985), it is useful to examine a particular idiolect with reference to its position on the cline of linguistic, stylistic, and social literacy.

2.2 Digital data sources for Scots

As early as the publication of his visionary article on variation and variety in Middle Scots in 1971, A. J. Aitken saw that new technologies would permit us to describe complex linguistic systems and to identify the factors conditioning the choice of variants. The pioneering work of Scottish lexicographers Aitken, Sir William Craigie and David Murison, along with the numerous experts who later joined the editorial staff of the *Dictionary of the Older Scottish Tongue* and the *Scottish National Dictionary*, is a major achievement, which will lend strength and momentum to a significant spread of interest in Scottish studies (see Dareau 2004, 2005; Kay & Mackay 2005). These two dictionaries are now freely available in the online *Dictionary of the Scots Language* (DSL, http://www.dsl.ac.uk/dsl/index.html).

As Williamson (2005) illustrates, the two major scholarly achievements in Scotland in the field of Scottish linguistics, the linguistic atlases of Older Scots, with the Edinburgh databases they draw on, and the dictionaries, complement each other in various highly relevant ways. Williamson compiled the online *Linguistic Atlas of Older Scots*, drawing on the *Edinburgh Corpus of Older Scots*, phase 1, c. 1380 to c. 1500. (http://www.lel.ed.ac.uk/ihd/laos1/laos1.html).

Meurman-Solin compiled the *Helsinki Corpus of Older Scots* (HCOS), 1450–1700, which is based on editions and comprises approximately 850,000 words of running text. The texts in the HCOS represent fifteen different genres: acts of Parliament, borough records, trial proceedings, handbooks, scientific treatises, pamphlets, sermons, the Bible, histories, biographies, diaries, travelogues, educational treatises, official letters and private letters (see Meurman-Solin 1993 and 1995). The HCOS is included in the CD-ROM containing the *ICAME Collection of English Language Corpora* (http://clu.uni.no/icame/newcd.htm).

See also http://www.helsinki.fi/varieng/CoRD/corpora/HCOS/index.html

A major resource for the study of Scots is the SCOTS corpus, the *Scottish Corpus of Texts and Speech*, available on the Internet (http://www.scottishcorpus.ac.uk/corpus). The team responsible for compiling this continually expanding database consists of members of the English Language Department and STELLA project of the School of English and Scottish Language and Literature, University of Glasgow. The project aims to create a large-scale electronic corpus of both written and spoken texts for the languages of Scotland (see also Anderson, Beavan & Kay 2007).

The *Helsinki Corpus of Scottish Correspondence* (ScotsCorr), which is introduced in the present manual, permits the wider international community to use manuscript sources which would otherwise be impossible to consult without visiting the various archives personally. In her pre-2006 publications, Meurman-Solin used the 2003-version of the CSC, using CorpusPresenter by Raymond Hickey as a search engine (Hickey 2000, 2003), whereas her publications in 2007–2012 draw on the CSC corpus (2007).

See also the digital data sources at the National Archives of Scotland (www.nas.gov.uk) and the National Library of Scotland (www.nls.uk), in particular.

2.3 Letters as a data source

This section aims to highlight aspects of letters which relate to their use as a data source in linguistic research. The points raised are based on the compiler's own research in the fields of

historical linguistics, historical sociolinguistics, stylistics, pragmatics, and dialectology, amongst others.

As amply evidenced by the *Corpus of Early English Correspondence* (CEEC; in 1996 2,4 million words) and its expanded versions, a database exclusively containing letters can be carefully structured according to socio-linguistically relevant variables such as the writer's social rank, gender, age, social and geographical mobility, and education (Nevalainen & Raumolin-Brunberg 1996, 2003). A large database, and a large number of informants, is required for a balanced account of social stratification. While the size of the CEEC is adequate for the application of the methodological principles and practices that have been specifically tailored to make it a valid tool for historical sociolinguistics, the smaller size of the ScotsCorr only permits the use of more general classifying criteria (for information on size, see Nurmi 2002).

Another distinctive feature of letters as a data source is related to their communicative function and circumstances of production. The data represents on-line language use in an explicitly interactive communicative situation. This means that there is relatively little or no editing. There is a rich variety of idiolectal grammars, some of them virtually unaffected by standardizing trends. Since some features of visual prosody in the manuscript texts (spacing, marked character shapes, etc.) have been digitized using an annotation system designed for this purpose, the on-line processing of thoughts in the interactive communicative act of addressing a recipient is recorded. Even hesitation, deletions, insertions, and corrections have been signalled as such in the digitized texts (Section 6 Visual prosody and Section 7 Symbols and Comments). Corrections of this kind can, of course, be interpreted as evidence of some degree of editing, but usually it is not possible to find evidence that a sequence of versions were prepared before the acceptance of the one to be sent to the addressee.

Even though letters remain mostly unedited, they do not necessarily represent unplanned discourse. This is because the form of a letter is quite strictly regulated by genre-specific schemata and conventions of epistolary discourse (e.g. Nevala 2004: 37-40, Barton & Hall 1999, Daybell 2001, 2012, Fitzmaurice 2002, Schneider 2005). In contrast with texts representing other genres in diachronic corpora, of which only a sample has usually been included, in correspondence the whole text of a letter can be examined. This is highly significant for research which adopts the semantic-pragmatic approach, and it also permits a detailed analysis of text structure with reference to discourse strategies. Earlier research has shown that it is particularly politeness strategies in general and formulaic language use which condition the choice of linguistic features (e.g., Meurman-Solin 2000c on the introduction of the relative *who* and Bergs 2005 on morphosyntactic variation in the Paston letters). For information on stylistic literacy as reflected in early correspondence, see Meurman-Solin (2001b) and Meurman-Solin & Nurmi (2004).

The ethnography of communication, including both situational aspects and those which define the participant relationship, can be reconstructed using both language-external factors and indirect evidence, the latter provided, for example, by the choice of terms of address and discourse strategies conveying respect. The role of letter-writing manuals will also have to be taken into account in interpreting the data (cf. Nevala 2004: 33-36, Sairio & Nevala 2013).

Unlike the *Helsinki Corpus of English Texts* (http://www.helsinki.fi/varieng/CoRD/corpora/HelsinkiCorpus/index.html and Rissanen & Tyrkkö 2013 http://www.helsinki.fi/varieng/journal/volumes/14/rissanen_tyrkko/) and the *Helsinki Corpus of Older Scots* (http://www.helsinki.fi/varieng/CoRD/corpora/HCOS/index.html), the ScotsCorr does not contain information about the variables of "level of formality" and "participant relationship" (see Kytö [1991] 1996 and Meurman-Solin 1993: 180-183). However, the user can

define the parameter values of these two variables by using the information provided about the writer/informant and the addressee at the beginning of each text (see the discussion of the parameters %IM/IF/IR (informant male/female/royal) and %AM/AF/AR (addressee male/female/royal) in Section 5 Language-external information in the text files).

Digital images of the letter manuscripts have not been provided in the present version of the ScotsCorr, and no attempt is made to describe the letters as physical objects (see, however, Meurman-Solin 2013a-c). However, the transcripts contain information about features such as the position of the date and place of writing, the positioning of the text on a folio or a number of folios, spacing and indentures, the use of margins, and any additional text positioned after the signature. With some exceptions (see Preface), the recipient's address on the folded sheet is also given whenever there is one on the original. It is obvious, however, that this information is insufficient for a full reconstruction of a letter as an object; for example, information about wax seals attached to letters is not provided.

3 Dimensions of space, time, and social milieu

This section provides general information about the variables of time, space, and social milieu for the user to consider in assessing the representativeness of the ScotsCorr corpus (for a very important discussion of these variables, see also Laing 2004).

Even though some basic language-external information is provided about each digitized letter (for a detailed account of the language-external information presented at the beginning of each text file, see Section 5 Language-external information in the text files), the main principle is to keep the texts and a knowledge bank in which information about the texts and their writers has been deposited separate. The rationale is that some of this knowledge is hypothetical in nature and may even prove incorrect later, as further research is available. With an increased understanding of a particular textual history, it may become necessary to rewrite its description. Thus the principle of proteanism (see Section 1.1.1 Protean corpora: multidimensionality, flexibility, and transparency) is applied to how the auxiliary databases attached to the ScotsCorr have been compiled and constructed.

As discussed in Meurman-Solin (2001a, 2003, 2004a), new users of corpora which have been carefully structured according to language-external variables may sometimes apply such variables as interpretative tools rather uncritically. For example, genre is often considered the primary factor not only in assessing representativeness but also in interpreting linguistic findings. Apart from the more general problem of it being 'difficult for the analyst to separate out the effects of diachrony from the effects of genre' (Herring *et al*. 2000: 3), there are problems caused by the fact that genres are often unevenly represented over time, space, and social milieu. 'Since there are gaps in historical data in the earlier periods in particular, the claim that genre is the primary conditioning factor in the introduction and spread of a specific linguistic feature is valid only in so far as the database the evidence is extracted from can be considered fully representative' (Meurman-Solin 2003: 172). Hopefully, a separate knowledge bank will motivate the users to reconceptualize and redefine these variables in accordance with both their theoretical and methodological approach and their particular research question.

In my view, the most important task at this stage of reconstructing and describing variation and change in Older Scots is to stress that conclusions should be data-driven and data-oriented. It would be unwise to interpret the findings with reference to sociolinguistically defined variables, for instance, before other factors – such as the conventions of epistolary discourse, partly borrowed or recontextualized from other discourses, or the influence of models, both British and European, on letter-writing – have also been thoroughly studied. In the genre of letters, the practices of polite society influence linguistic and stylistic preferences in an important way (Palander-Collin 1999, Nevala 2004, Meurman-Solin & Nurmi 2004, Palander-Collin & Nevala 2005, Sairio & Nevala 2013).

3.1 Time

The very earliest Scottish letters in the archives date from circa 1400 (those by George March in 1400 and James Douglas in 1405). Despite continued browsing through the family archives kept in libraries, record offices, and private collections, the size of the part of the corpus dating from the fifteenth century, and the proportion of autographs in particular, will remain quite small. In a corpus of letters, a certain degree of imbalance between the quantity and quality of pre-Reformation evidence and data from later periods is unavoidable (a similar problem occurs in the CEEC; see Nevalainen & Raumolin-Brunberg 1996, 2003, Nurmi 2002), while the situation is quite different as regards early legal and administrative documents and literary texts. These form the core of the the *Linguistic Atlas of Older Scots* (LAOS), compiled by Keith Williamson, phase 1, c. 1380 to c. 1500 (http://www.lel.ed.ac.uk/ihd/laos1/laos1.html).

The most valuable data in the mid-sixteenth century can be extracted from deposit SP2 in the National Records of Scotland (NRS), which contains letters written by geographically diverse correspondents to Mary of Lorraine, Queen Dowager (widow of James V). A significant proportion of these letters are autographs and, since they date from a relatively short period (1542–1560), they also provide useful evidence for a synchronic study of diatopic and diastratic variation (Meurman-Solin 2000a). As regards pre-1560 texts, it has not been possible to achieve this degree of representativeness in other twenty-year periods (the periodization adopted in the *Corpus of Early English Correspondence*; see Nevalainen & Raumolin-Brunberg 1996). In contrast, there is no lack of manuscript sources in the genre of correspondence in the later periods. The ScotsCorr is designed to be diachronically representative up to 1750, the seventeenth century being the most densely covered period in the present version of the corpus. A supplementary, although somewhat different, corpus of Scottish correspondence is being compiled by Marina Dossena, University of Bergamo (Dossena 2004, 2012, 2013, Dossena & Del Lungo Camiciotti 2012). This corpus chiefly contains nineteenth-century business letters and letters written by emigrants. For information about the variable of time in the text files (the parameter %DA), see Section 5 Language-external information in the text files.

For further information on the variable of “time”, see Section 2.1.

3.2 Space

The following fifteen regions are represented in the present version of the corpus: Aberdeenshire, Angus, Argyllshire, Ayrshire, Border counties, Fife, Invernessshire, Lanarkshire, Lothian, Moray,

Perthshire, Ross, South-West (Dumfries and Galloway), Stirlingshire, and Sutherland. Neither the Orkney and Shetland Islands nor the Western Isles have been included. The database permits the reconstruction of the dialect continuum in Scots, especially for the seventeenth and the latter half of the sixteenth centuries.
The above-listed fifteen regions can be grouped to represent the areas of North (25 per cent), North-East (14 per cent), Central (12.5 per cent), South East (35.5 per cent), and South-West (13 per cent of the informants) (for an example of statistical analyses based on these five larger regions, see Meurman-Solin 2012).

Region	Male	Female	Total	%
North	72,775	19,151	91,926	24.9
North-East	40,519	11,241	51,760	14.0
Central	31,708	14,334	46,042	12.5
South-East	99,819	31,123	130,942	35.5
South-West	43,533	4,317	47,850	13.0
Total	288,354	80,166	368,520	100.0

Geographically defined informants in the Helsinki Corpus of Scottish Correspondence 1540-1750.

Of the altogether 466 informants, 43 remain unlocalized (c. 15,000 words; c. 4 per cent). It is noteworthy that there are two categories which have not been defined by the geographical origin of the writer: the parameter value "Professional" has been given to lawyers and members of the army and the clergy (c. 28,000 words; c. 7 per cent) and the value "Court" to a small sample of letters by Queen Mary, James VI, and three Regents of Scotland (c. 5,000 words). The total number of letters is 1,362. See the auxiliary file ScotsCorr Quantitative Data.

The place of origin of the writers and, if specified in the original, the place of writing of a given letter are indicated as part of the set of file-initial parameters as well as in the various auxiliary files. The scarcity of prosopographical information about some informants, women in particular but also men representing lower social ranks, may make their localization by language-external factors difficult. As a result of marriage (or marriages), women may have moved from one place to another. Moreover, since we do not usually know where and when a female informant became literate, it may be impossible to tell which area the spelling practices she learned belong to and how they relate to her pronunciation. The proportion of female informants in the corpus is approximately 21 per cent (more precisely, 19.2 localized and 1.5 unlocalized).

If an informant cannot be localized using language-external criteria, it may be possible to position his or her idiolect in a particular geographical area according to linguistic criteria, by applying the principles of the "fit technique" (see Laing & Williamson 2004, for instance). For information about how the variable of space is integrated in the text files (the variable %LC), see Section 5 Language-external information in the text files.

For further information on the variable of "space", see Section 2.1.

3.3. Social milieu

The user may decide to interpret linguistic findings with reference to various factors, including the writer's and the addressee's gender, age, social rank, social mobility, and education. The user may also try to reconstruct the informants' social networks, including the geographical spread of such networking. However, in the case of numerous women and younger sons in particular, this is only possible to the extent that such information is provided by their correspondence.

The age range is relatively representative, with the exception of very young informants. Patrick Waus' letters to his parents, written when he was of school age (circa 1540), have not been included, since it has not been possible to check them against the manuscripts. The fate of the Waus Correspondence is unknown; according to the NRS, these letters may have been destroyed in the fire at Barnbarroch House in 1941. The user may find it interesting to examine the editions of Patrick Waus' letters in the *Helsinki Corpus of Older Scots* (see also Meurman-Solin 1999).

The auxiliary databases Male Informants in the Helsinki Corpus of Scottish Correspondence, Female Informants in the Helsinki Corpus of Scottish Correspondence, and Royal Informants in the Helsinki Corpus of Scottish Correspondence provide language-external information which is based on standard reference works such as the *Scottish Peerage* and memoirs of some of the most renowned Scottish families, collected and edited by Sir William Fraser, as well as drawing on the continuously improving catalogues of the libraries and archives in Scotland. Since the inclusion of more recent historical, sociological, and genealogical research would require an interdisciplinary team of researchers, this auxiliary resource remains incomplete and is not available internationally. Since sources on historical research of this kind were not available online when this work on Scottish correspondence began, the compiler humbly admits that information about the writers and the addressees is often insufficient. The user of the corpus is advised to consult the continuously expanding and improving online sources of information made available in more recent years.

My earlier studies drawing on corpora of Older Scots have shown that, in general, no straightforward correlation between linguistic variation and sociolinguistically-defined conditioning factors is evident. The spread of the relative pronoun WHO in sixteenth-century Scots reflects developments in how it is used as a reference signal, not only in noun phrase structures but also as a sentence-level constituent. Since the early instances are frequently attested in formulae (typically the final formula '[as knows] god who keep/preserve [the addressee of the letter] eternally'), the history of WHO in Scots can be related to the spread of stylistic literacy (Meurman-Solin 2000c). Early Scottish women's writing skills have been illustrated in Meurman-Solin (2001b), and social milieu rather than formal education explains the higher degree of stylistic literacy of some of the female informants. Meurman-Solin & Nurmi (2004) examines the use of circumstantial adverbial clauses introduced by *seeing* and *considering*. These topic-forming clauses are skilfully used by numerous letter-writers to provide background information of various kinds (see also Meurman-Solin & Pahta 2006). Meurman-Solin (2002) shows that the progressive is more frequent in two specific environments, in which its use can be shown to be conditioned by text type: these are narratives and speech-based texts, depositions of witnesses in trial proceedings being examples of the latter. Thus, the frequencies and distributions of particular linguistic features can be related to various discourse properties. Johnston (1997: 51) draws our attention to a more general concern related to stylistic competence, claiming that more mobile people would use a "watered-down" style to communicate with people from other regions; apart from this better ability to differentiate between speech styles, the upper classes and professionals, in general the more mobile people, would act as the main guardians of a "Standard Scots style". By this logic, 'the town vernacular might well be different from the countryside ones just outside the walls, given that cities did tend to

have a more varied population and would attract people from around their whole hinterland, which could extend over more than one dialect group.'

The user may find it appropriate to redefine the concept of space, extending it to cover dimensions other than geographical area. My earlier corpus-based research suggests that "distance" as a social, economic and cultural construct, rather than as a concept defined purely geographically, is a significant conditioning factor in the variation and change attested in the history of Scots (Meurman-Solin 1999, 2000a-c, 2001a).

In the present approach, diastratic variation can be said to complement our understanding of the spread of linguistic features over time and space. Thus, the primary goal of my research has been to create variationist typologies of linguistic systems by drawing on minutely detailed inventories of data. In this approach, membership of a variationist typology is strictly limited to items that have been attested as genuine alternatives in a pattern of variation at a particular level of analysis, whether structural, syntactic, or related to communicative or text-structuring functions.

3.4 Community type

In addition to time, space, and social stratification, the representativeness of a database can be assessed with reference to three community types: "speech community", "discourse community", and "text community". As suggested in Meurman-Solin (2004c: 28), 'the best informants for reconstructing practices of a speech community can be found in texts written in private settings by non-professional, preferably less trained and relatively inexperienced writers.' Phonetic spellings and other features reflecting spoken varieties recorded in letters by these informants are not attested in texts influenced by shared scribal practices or, in the case of early printed works not available in manuscript, by the preferences of printers. For information on recorded phonetic spellings, see Meurman-Solin (1999, 2001b and 2005).

As argued in Meurman-Solin (2004c), while letters written by informants defined as representatives of speech communities permit us to identify idiolectal grammars, those by members of discourse communities also reflect grouplectal preferences. Language use in texts of the latter kind is affected by the conventionalised practices of professional coalitions, writers sharing similar communicative goals and applying similar genre-specific rules of writing, or groups who strictly follow a specific prescriptivist trend. Texts created by members of a particular discourse community can no longer be exclusively examined with reference to the variables of time, space, and social milieu, since at least some of their linguistic choices have been influenced by 'inherited, borrowed, or recontextualized discourses, English or foreign' (Meurman-Solin 2004c: 28).

The term "text community" refers to literate people in a particular place and time who share a particular range of written texts. The identification of a text community is based on information about the consumption of literary texts and texts representing religious instruction. Another method for reconstructing text communities is to browse through bundles of documents put together in the archive of a particular family, administrative body, or some other institution. Such bundles will typically contain legal documents and letters to officials or friends and relatives, but also more or less unedited reports, notes, *pro memoria*-type documents, diaries and memoirs. Text communities have tended to be defined on the basis of edited texts, and many texts, despite their integral social and communicative function in their historical context, have tended to be marginalized, as they do not have the status of texts or genres traditionally included in the canon.

The conceptual framework provided by the three community types permits us to assess the relevance and validity of databases more reliably. We will become aware of the extent to which the range of texts varies between communities, for instance. For example, sixteenth-century Scottish women, 'mainly used their writing skills for writing letters to their relatives, and, somewhat later, for keeping accounts and summarizing the daily events in their personal diaries. In this case, language use can be assumed to be essentially conditioned by the restricted social functions of writing' (Meurman-Solin 2001a: 16). For information on Scottish women's literacy and education, see Marshall (1983) and Houston (1985).

A diachronic database representing a text community would in theory comprise the full range of once-functional texts relevant to the expression of the communicative purposes of the various discourse communities in a given geographical area. In practice, this full range will remain beyond recovery and it is therefore necessary to provide at least some direct or indirect evidence of what the major gaps are.

4 The ScotsCorr database

4.1 Selection of data for the ScotsCorr

The manuscripts of the letters included in the *Helsinki Corpus of Scottish Correspondence* are deposited in the National Records of Scotland and the National Library of Scotland. References to the catalogues of these archives are given at the beginning of each document (for information on the references, see the discussion of the symbol %MS in Section 5 Language-external information in the text files). Since the compiler lives outside Britain, it has been necessary to restrict the focus of the corpus to these major collections of Scottish letters. Inevitably, the present version of the ScotsCorr only contains a small proportion of these important collections. For a corpus of this size, it was unnecessary to search for more material by visiting the various local archives. Of course, some early correspondence is still in private hands, and less accessible than the family archives deposited in public record offices. As regards material to which access is limited, the compiler has contacted the trustees of such documents, and been granted permission to make her transcripts of the documents available as part of the ScotsCorr database distributed to the academic community internationally.

The user should keep in mind that the compiler's copyright only applies to the annotated transcripts of the letter manuscripts. The copyright of the source manuscripts lies with their repository or, in some cases, with private owners. The original manuscripts used to compile the ScotsCorr are deposited in the National Records of Scotland and the National Library of Scotland, Edinburgh, UK, and, as regards the right to reproduce or to publish the manuscript sources of the historical documents themselves, the user should contact these institutions for more information.

The present ScotsCorr has been designed to provide as much information about late sixteenth-, seventeenth-, and early eighteenth-century correspondence as possible. The very earliest Scottish letters extant in the archives date from circa 1400 (see Section 3.1). Since these are included in the LAOS database, which is an important manuscript-based source for Scottish documents dating from the fifteenth century compiled by Keith Williamson, the ScotsCorr is restricted to post-1540 letters.

Despite continued browsing through the archives, the proportion of fifteenth-century letters in the corpus will remain small. While, as a result of much more work at the archives, the number of fifteenth-century autograph letters cannot be expected to increase to a considerable extent, there is no lack of sources in the genre of correspondence for the later periods.

The present version of the corpus also comprises a selection of letters dating from the first half of the eighteenth century. The focus on the early part of the eighteenth century is primarily due to the fact that later letters reflect a considerable widening of contacts with English writers, and it seemed important to examine epistolary prose in primarily Scottish networks before extending the corpus to include letters by informants regularly commuting between Scotland and England.

The designation of a writer as of Scottish origin is based on information available in biographical sources of various kinds. Thus, Scottishness is exclusively defined by language-external criteria, the family background being a primary source of information in the categories of the nobility and the gentry. It has not been possible to provide conclusive evidence of the informants' geographical mobility. Some writers (43 informants; c. 15,000 words) in the database remain unlocalized.

The balance between the sixteenth and the seventeenth century has not yet been fully achieved: the proportion of sixteenth-century letters is smaller and consists of fewer informants.

	1540-1599	1600-1649	1650-1699	1700-1749	Total	%	N Informants	N Letters
Male	45,514	159,759	79,429	41,474	326,176	78.1	340	1,000
Female	2,468	32,113	37,041	14,699	86,321	20.7	118	335
Court	3,669	1,543	-	-	5,212	1.2	8	27
Total	51,651	193,415	116,470	56,173	417,709		466	1,362
%	12.4	46.3	27.9	13.4		100.0		

Informants in the Helsinki Corpus of Scottish Correspondence 1540-1750.

The most valuable data dating from the mid-sixteenth century have been extracted from a collection catalogued as SP2 in the National Records of Scotland (NRS), which contains letters written by geographically diverse correspondents to Mary of Lorraine, Queen Dowager (widow of James V). These date from a relatively short period 1542–1560. Only the letters identified as autograph in the earlier edition of these letters by Annie I. Cameron have been included.

The main criteria in the selection of data for the ScotsCorr corpus are as follows:

Only original manuscripts of letters have been included; there are no letters which have been indicated to be later copies in the catalogues or the actual documents, or have been detected to be such by the compiler according to criteria such as type of handwriting and paper quality.

Priority has been given to autograph letters by a single writer, those by two or more writers being exceptions in the ScotsCorr (see Index of Sources). However, it has not always been possible to find conclusive evidence for a particular letter being by the hand of the person who signed that letter. Comparison of hands is not always possible, since the archives have had to limit the number of documents a reader is allowed to examine simultaneously. There may be other reasons that a hand remains unidentified; for example, we may have only one single letter in a particular hand.

Among the sixteenth-century letters in particular, there are letters written by two different hands. In these, the most frequent pattern is that the body of the letter is in secretary hand and the signature, sometimes also the letter-closing formula, and, even less frequently, the initial term of address, are in a different hand, mostly resembling italic or a variety of the more rounded styles. In letters of this kind, the section in secretary is assumed to be non-autograph, while the signature (and the formulae) are considered autograph. The two hands are indicated by positioning the comment {hand 1>} before and {<hand 1} after the autograph sections and {hand 2>} before and {<hand 2} after the non-autograph ones. However, it is necessary to keep in mind that the script type or style of writing in the signature may be different from that in the body of the letter, but this difference does not always indicate that these are by two different writers. For information on the file-initial commentary, see %HD1 and %HD2 in Section 5 Language-external information in the text files.

As discussed in Section 3 Dimensions of space, time, and social milieu, while the chief goal has been to achieve diachronic, diatopic and diastratic representativeness, close attention has also been paid to ensuring that the proportion of letters written by and addressed to women does not remain too small.

The selection process has also been conditioned by a number of pragmatic issues. There may be a tendency to include more numerous documents from carefully catalogued compilations which are easy to access at the archives. A particularly important factor in the decision-making has been the physical condition of the documents. Badly damaged documents have usually been excluded, especially those in which the folio is either torn or worn out at the margins, or where it has stayed folded for centuries. Since the keepers of the documents have been obliged to disallow their reproduction by photocopying, some of these documents, as well as those in which the ink is very pale, have been transcribed *in situ*. The transcriptions of these letters have usually been rechecked during another visit to the archives. However, the majority of the ScotsCorr letters have been transcribed from photocopies or photographs ordered by the compiler. In the case of imperfect copies, the originals have been re-consulted. For information about the transcription process and the "CSC archive" containing the copies, see %ST in Section 5 Language-external information in the text files.

Sample size, indicated by the number of words (%WC), varies between the different informants for two reasons. Letters differ from one another considerably as regards their length; for example, letters by legal or financial advisers are usually much longer than the rather formal letters written by newly-married women to new relatives as a polite gesture. Since it has not been possible to regulate sample size, frequencies will have to be normalized in the presentation of the results of quantitative analysis. Word counts are provided in %WC at the beginning of each file. Statistical information in the auxiliary file Wordcounts by Individual and Locality also contains word counts for all the letters by a particular writer, as well as providing totals for all letters localized to a particular region. However, it should be remembered that letters representing a particular geographical area may sometimes be linguistically too heterogeneous to permit the interpretation of the findings with reference to dialectal preferences.

When the categories of "Professional" and "Unlocalised" are kept distinct from the majority of male and female writers, the word counts in the four periods are as follows:

	1540-1599	1600-1649	1650-1699	1700-1749	Total	%	N Informants	N Letters
Male	37,501	146,334	72,029	32,490	288,354	69.0	289	915
Male Profess-ional	5,293	10,847	4,826	7,250	28,216	6.8	23	55
Male Unlocal-ised	2,720	2,578	2,574	1,734	9,606	2.3	28	30
Female	2,190	30,167	34,264	13,545	80,166	19.2	103	313
Female Unlocal-ised	278	1,946	2,777	1,154	6,155	1.5	15	22
Court	3,669	1,543	-	-	5,212	1.2	8	27
Total	51,651	193,415	116,470	56,173	417,709	100.0	466	1,362
%	12.4	46.3	27.9	13.4	100.0			

Informants in the Helsinki Corpus of Scottish Correspondence 1540-1750: geographically defined, professional, unlocalised and royal.

4.2 Transcription and digitization

There have been previous editions of Scottish correspondence, as part of the Scottish History Society publications (e.g., *The Scottish Correspondence of Mary of Lorraine, 1542-1560*). A major achievement in the field is the series of volumes compiled by Sir William Fraser in the nineteenth century. These contain the memoirs of some well-known Scottish families, as well as charters and letters directly related to their history (e.g., *Memorials of the Montgomeries, Earls of Eglinton*, 2 vols, Edinburgh 1859; *The Chiefs of Grant*, 3 vols, Edinburgh 1883; *Memorials of the Earls of Haddington*, 2 vols, Edinburgh 1889; *The Sutherland Book*, 3 vols, Edinburgh 1892). These editions seem to have been produced for an intended readership consisting mostly of historians. Even though in principle the language of the documents has not been modernized, the introduction of modern punctuation and sentence structure, for example, prevents linguists from using these editions as a valid data source for studies of syntax, discourse and text structure in particular. The differences between a transcript of a letter prepared for the ScotsCorr and the edition of the same letter in *Memorials of the Montgomeries* by Sir William Fraser have been illustrated in Meurman-Solin (2001: 20-21). A more detailed study is available online (Meurman-Solin 2013a).

Another major problem in the nineteenth- and early twentieth-century editions is caused by the practice of tacitly expanding contracted forms, often replacing them with a full form which has been selected without giving a justification for the preference of one variant over another. The most serious problem is the fact that, in general, the editors provide very little information about their editorial principles and practices; for historical linguistics, this information is clearly insufficient.

Since the Scottish language dictionaries such as the *Dictionary of the Older Scottish Tongue* are also based exclusively on edited sources, it has been difficult to find valid data for the

reconstruction of the history of the Scots language. The *Helsinki Corpus of Older Scots* (HCOS) (http://www.helsinki.fi/varieng/CoRD/corpora/HCOS/index.html) is also based on editions. As a compiler and user of the HCOS corpus, I have found it necessary to keep in mind that, in such areas of research as phonology and partly also morpho-syntax, findings based on this corpus may reflect the history of varying editorial principles and practices rather than the history of the Scots language.

In contrast, the *Helsinki Corpus of Scottish Correspondence* (ScotsCorr) applies the principles and practices of philological computing to the transcription and digitization of the manuscript originals. In other words, the text in the original manuscript has been reproduced faithfully, and emendation, tacit expansion of contracted forms, modernization and normalization have not been permitted. Question marks signal ambiguous readings, and indication is given of cancellations, insertions, and non-linguistic features, i.e. visual prosody, in the manuscript originals. The various practices are described below, while the annotation of visual prosody is discussed in Section 6 Visual prosody.

4.3 Features requiring comments or annotation in the ScotsCorr letters

Section 4.3 chiefly discusses the transcriber/digitizer's comments on the manuscripts and such features in them that affect decisions made about transcription and digitization practices, whereas Section 7 Symbols and Comments in the ScotsCorr provides a detailed description of the symbols used by the compiler/transcriber to signal out particular linguistic and script-related features of the manuscript texts and the typology of her commentary.

4.3.1 *Manuscript layout*

No detailed description of where on a folio or folios a text has been written is provided, nor are other features, such as quality of writing material, quill, or ink, described; comments are restricted to features that may be relevant in the analysis of linguistic features. However, the number of folios (e.g. {f1} and {f2}) is given, as well as information about the position of a text on a single folio ({f1r} and {f1v}). Unfortunately, since the compiler has had to rely on photocopies of the original documents, a distinction is not always made between text on two folios and text on both sides of a single folio; it has not been possible for the compiler to recheck manuscripts that were seen and copied at some earlier stage. This distinction does not seem to play a significant role in linguistic research, so, for reasons of economy, a policy was adopted to focus on features that can be claimed to be relevant, in this case whether the text of a letter continues on a new page, be it the reverse side of the same folio or a new folio. It is also quite common for the text to be written in two columns on a page, usually starting from the right side and continuing on the left. This is also signalled by {f1r} and {f1v}. Thus, because of lack of information about decisions made in the copying of the documents, {f1r} may refer to the right side of a folio or the first folio of a multi-page letter, and {f1v} the left side of a folio or its reverse side.

Two types of comments are used for indicating text written in the margins, almost always the left one. If a letter continues in the margin, this is marked by {in margin>}. In cases of this kind, the direction of writing usually changes; this has been indicated by {direction changes>}. A note or a correction in the margin is treated as an insertion, and marked by being bracketed using the pair {ins} … {ins} and {<in margin}.

Both line-breaks, marked by \ (i.e., a backward slash), and paragraph structure, marked by \\ (i.e., a double backward slash), are indicated. Paragraph structure can also be reconstructed by searching

for the comment {left indenture>} or features of spacing ({space}, {a space vertically}, or {a wide space vertically}. As regards features such as the date and place of writing, the initial term of address, the pre-signature formula and the signature, and post-signature insertions, the layout can be reconstructed from the position of these in the running text (for example, the date and place of writing may precede the body of the letter rather than occurring after the signature) and by the above-mentioned system of backward slashes. At the end of lines, both <-> and <~> occur in the manuscripts, being added by the writer either just to mark an empty space or to signal text structure, i.e., the end of a chunk of discourse and the beginning of the next.

The text which functions as the recipient's address is usually written on the folded folio, the form usually being the preposition *to* or *for* and the addressee's title and name (e.g., *To the Countess of Findlater*), sometimes followed by *These* ('Deliver these to the Countess of Findlater'). Only rarely do we find a reference to the addressee's dwelling-place. In the transcript, these addresses are preceded by the comment {address>}.

These features are discussed and illustrated in Meurman-Solin 2013b. For detailed information, see Section 7.

4.3.2 *Insertions, deletions, cancellations, and corrections*

Greetings from the writer's family members are often added after the signature (for further information, see Meurman-Solin 2013b). These – as well as any other additions that come after the signature – have been considered to be insertions, and are therefore bracketed by {ins} comments. The same pair of comments is used with all insertions, whether these are inserted characters, words, or longer chunks of text. Deleted full words or longer chunks of text are bracketed by {del} comments, but, despite being cancelled by the writer, these items occur in the Korp concordances. A deletion followed by a correction is marked as follows: {del} I {del} {ins} we {ins}. When an insertion or a deletion occurs word-finally, a single {ins} or {del} is used (e.g., servan{ins}t).

However, if a deleted word can no longer be read, because of thick strikethroughs, for instance, the comment {cancellation} is used. This comment is also used with deletions which do not constitute a full word. This is the case with false starts consisting of a character or two, or in cases where a character or characters have been crossed out as a correction, for instance. In other words, the comment {del} … {del} occurs with features included as words in concordances, whereas the comment {cancellation} is used with illegible or fragment-like features.

When a word can no longer be read because of damage to some of the characters (for example, when only the characters <al> remain, a comment such as {<torn except for initial/medial/final <al>} replaces the damaged word in the running text. Thus, instead of an emendation, incomplete words contain the symbol # to indicate where the missing part is in a word (e.g., al#). In some unambiguous cases in which the context permits only a single reading, a question-mark {?} is added to each character no longer clearly visible in the manuscript. This practice seems appropriate especially in cases in which the editor of a previous edition appears to have seen the missing characters. When the original manuscripts are deposited in the archives, they are sometimes inserted – even using glue – in bound volumes without taking proper care of the folios remaining fully visible. This often explains why what a previous editor has seen differs from what is visible in the manuscript or its photocopy available to the ScotsCorr compiler.

A correction may remain illegible or cause ambiguity. A question-mark immediately following an unclear character indicates problems of this kind, comments such as {<an unclear correction} or {<corrected} explaining its use.

4.3.3 *Ambiguity*

Each transcript has been checked against the manuscript at least twice, or, in the case of letters with copies in the ScotsCorr archive, at least three times. The number of irregular or untidy hands is quite large in the genre of letters, many of them having been written by untrained and inexperienced writers. This is sometimes reflected in ambiguous realisations of characters. Some letters are hurriedly written notes; some hands are just untidy. It is therefore not always possible to suggest one particular reading. The following practices have been used to help the user spot these ambiguities.

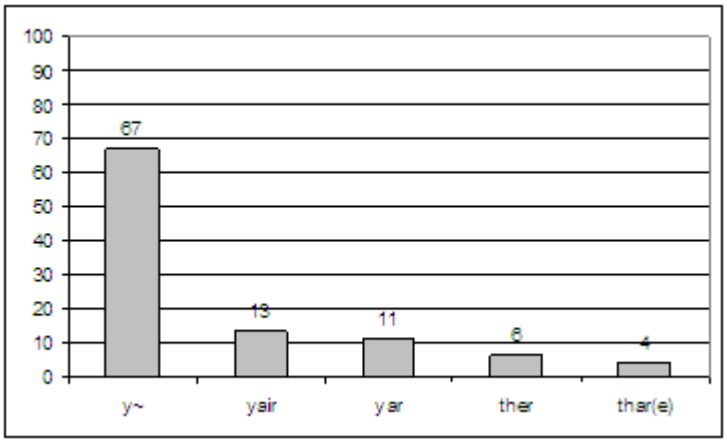
A comment positioned before the body of the letter may provide information about a particular set of characters which are consistently difficult to distinguish. Sporadic instances of ambiguous characters are followed by a question-mark (e.g. *condic?ioune*, in which <c?> indicates that the reading could also be *conditioune*, the shapes of <c> and <t> being close to one another in some hands). As is often the case with <c> and <t>, some other doubtful readings may permit the suggestion of two alternatives. For example, the compiler/transcriber may suggest *sa?me* and add a comment {<or <o>} to indicate ambiguity between <a> and <o> in this particular instance. Similarly, in the above-mentioned case of ambiguity between <c> and <t> the comment {<or <t>} follows the word *condic?ioune*.

Ambiguity resulting from physical damage is indicated with {<blurred}, {<torn} or {<damaged} positioned after the ambiguous word. When a hole in the original manuscript hinders the reading of a particular word-form, two question-marks indicate that there is space for one character and three question-marks that there is space for two or more characters. Similarly, two question-marks signal that a particular character cannot be read, and three are used to signal a longer sequence of unclear or damaged characters. Since the folio is often filled without leaving a margin, and the manuscript folios are often damaged at the edges, ambiguous word-final characters are particularly frequent at the end of lines.

The realisation of characters may be careless, which may reflect either an idiosyncratic tendency in a particular hand or a situational factor such as the letter being written in haste or there being insufficient space available. Another idiosyncratic practice that has been recorded is sequences of incompletely realised and often merged characters; these are indicated with either {<compressed} or {<reduced}. See Section 7 Symbols and Comments in the ScotsCorr.

4.3.4 *Contracted forms*

One of the most widespread practices in earlier editions is for contracted word-forms to be expanded surreptitiously. This considerably reduces their validity and relevance as data. However, only a relatively limited set of lexical items have contracted variant forms in the ScotsCorr data, and their frequency decreases by the third quarter of the seventeenth century. In fact, contracted forms chiefly occur in letters written in secretary hand. The general frequency of contracted forms of particular words can be illustrated with the following figure, which is based on the 2003 version of the CSC (see Meurman-Solin 2004c: 30). Of the 102 occurrences of *there* in adverbial compounds such as *thereof* and *thereafter* in the *Correspondence of Mary of Lorraine*, the proportion of contracted forms (here realised by y~) is very high:



The editorial practice applied in expanding *y~* in this 1927 edition by Annie I. Cameron remains opaque. In this edition as many as 67 per cent of the occurrences turn out to be unreliable evidence for the reconstruction of the pattern of variation in the use of this particular item. For further information on differences between the nineteenth- and early twentieth-century editions and the ScotsCorr transcripts, see Meurman-Solin 2013a.

The main principle in the ScotsCorr is that lower case and upper case are used exactly as in the original manuscript. Expansions of contracted word-forms are explicitly indicated, and 'the contracted part' is put between the symbols *…%. As regards 'the contracted part', in the original manuscript the contraction is usually marked by a flourish of some kind, a particular character shape (in *per% and *con%, for example), or a line above or at the end of the contracted word. For example, the contracted variant of *your* is frequently *yo~* in the manuscript, and this variant is transcribed as *yo*ur*%. The element *ur % is used as an emic representation of all the possible variant realisations that the flourish could be a 'substitute' for. In Older Scots the variant *yowr* is also quite frequent. The use of a fixed representation is necessary for tailored searches, but the user will have to keep the full word-forms and the contracted forms distinct, never forgetting to make a distinction between them in their analysis of linguistic findings. There is no linguistic justification for selecting a particular expansion rather than some other variant; the choice is merely the transcriber/digitizer's decision, based on such pragmatic concerns as retrievability.

Abbreviated word-forms – distinct from contracted forms by there being no flourish – have not been expanded, but marks indicating a clipped form in the original manuscript have been retained. Thus *Lo* for *lordship* is usually written . Lo . or / Lo / in the original. If this word, for example, is written using just the initial L, the comment {=lordship} follows.

A pair of the symbol = indicates that the character or characters between these symbols is in superscript in the original manuscript (e.g., kny^{t} 'knight' is digitized as kny=t=). These word-forms are not expanded. Meurman-Solin (1993) examines variation between variants such as *richt* and *right*, drawing on data in the edition-based *Helsinki Corpus of Older Scots* (HCOS). In preparing the manuscript-based ScotsCorr, it has become obvious that these words are often contracted in the manuscript originals (*ry=t=* being particularly frequent), and that a statistical account based on editions is not valid.

4.3.5 *Character shapes and special characters*

A file may contain comments on idiosyncratic character shapes. These are positioned after the list of language-external parameters (see Section 5 Language-external information in the text files) and

immediately before the text of the letter. The function of this information is to describe the principles and practices applied to the transcription of particular character shapes in a particular idiolect. A comment aims to make the compiler's decision making transparent:

{a superscript <r> has been chosen, instead of a contraction, since the shape of <r> in superscript is the same as in various other positions}

The function of this comment is to alert the user to the compiler's policy of distinguishing between the shapes of <r> in full words and in superscript and the shape which acts as a flourish in contracted word-forms.

Another comment may point out that the compiler considers a particular character shape ambiguous:

{word-final <s> sometimes resembles the flourish representing a contracted plural morpheme transcribed as *is%}

This comment provides information about the ambiguity between the shape of <s> and the loop-like character functioning as a flourish for the plural morpheme *is% (e.g., *news* and *new*is*%). It has not been possible to give a detailed description of each hand; in order to make the corpus an even more reliable data source, the publication of digital images of the manuscripts would be required.

Some characters do not retain their original shape in the digitized texts. Thus, thorn is rendered by <y> (e.g., *þir* by *yir*), yogh by z (e.g., *ȝeir* by *zeir*), and the character ß by s*s% (e.g., *expenß* by *expens*s%). The compound characters <æ> and <œ> have been replaced by <ae> and <oe>. In the sixteenth and early seventeenth centuries the character <j> is infrequent, being usually written by using <I>. This capital <I> is retained in the transcript; however, when a dot appears above this capital <I>, the comment {<with a dot} has been added. Also <t> without a horizontal stroke is commented on by adding {<without a horizontal stroke} when it occurs in words in which this may cause ambiguity.

Accent-like features have been omitted. The most frequent among these is the symbol above <u>, which is used to mark the distinction between <u> and other minims.

Character shapes of a particular kind have also been depicted by comments such as {<<…> enlarged} and {<<…> extended}. The former describes the size, often also the shape, of a character as clearly larger than that of the same character in upper case elsewhere in the text (e.g., *My* {<<M> enlarged) *Dear Lord* in a term of address), whereas the latter comments on the shape, often also the size, of a character which is clearly extended, or stretched out to cover more space, in comparison with the same lower- or upper-case character elsewhere in the text (e.g., *And* {<<a> extended} at the beginning of a new chunk of discourse). Both comments provide quite relevant information for reconstructing syntactic and discourse structure in texts in which the use of punctuation and capitalization has not become regularized (Meurman-Solin 2013a).

4.3.6 *Punctuation*

Punctuation, if there is any, remains as in the original. However, it has not been possible to find a digital representation of all the various shapes a virgule may have in the letters. In earlier letters, a punctuation mark resembling the shape of a forward slash is positioned between words, with a space preceding and following, whereas in later texts its position lowers down until it is clearly below the line and is written closer to the preceding word. The slash-like shape and its numerous

variants in the transition period have been realised using the sign /; the shape resembling a modern comma has been digitized as ,. Other punctuation marks have been rendered by their modern equivalents. For more information, see Meurman-Solin 2013a: 4.2).

5 Language-external information in the text files

Language-external information about the letters and their writers and addressees is provided at the beginning of each file (each letter is a separate file in the ScotsCorr corpus). This information is structured according to a set of parameters. The same information is available in the auxiliary databases Male Informants in the Helsinki Corpus of Scottish Correspondence, Female Informants in the Helsinki Corpus of Scottish Correspondence, and Royal Informants in the Helsinki Corpus of Scottish Correspondence.

The file-initial information is structured as illustrated by the following example:

```
# 149
%MS: NRS GD3/5/49
613
0 0 0
+Perthshire+
+unspecified+
=DrummondRJean=
%ST: a copy in the CSC archive
%DA: 1613 March 17
%CO: by Jean Drummond, Countess of Roxburgh, to Anna Livingston, Countess of Eglinton
%BI: previously edited by Sir William Fraser in the Memorials of the Montgomeries, vol. 1,
33:190-191
%IF: Jean Drummond, Countess of Roxburgh
%AF: Anna Livingston, Countess of Eglinton
%HD1: autograph, italic
%LC: Perthshire, unspecified
%FN: DrummondRJean6130317
%WC: 328
{CSC 2007}
```

Each file begins with the symbol # followed by an identification number (149 in the above example). The various parameters providing language-external information are introduced by the percentage sign and marked with a symbol consisting of two upper-case characters and a colon. In addition, the parameter HD ‘hand’ categorizes autograph by 1 and non-autograph by 2.

%MS: stands for ‘manuscript’ and gives the reference of each document as catalogued in the archives. NRS is the acronym for the National Records of Scotland and NLS for the National Library of Scotland. GD refers to the Gifts and Deposits kept at the NRS, whereas Adv is an abridged form of the Advocates’ Library and Dep of Deposits at the NLS. MS occurs as part of the parameter value only if it is also part of the reference in the catalogues.

This information is followed by the year of writing, given as a three-digit number (686 for 1686). The zeros on the following line may later be replaced by coordinates based on the Ordnance Survey maps, which will allow the production of digital maps, permitting the presentation of data extracted from the ScotsCorr in the format of a linguistic atlas. The LAOS database applies this system of coordinates, so that, by also adding these coordinates to the ScotsCorr, the two corpora could be used as a combined data source in the future. However, the localization of all the ScotsCorr informants and the addition of coordinates based on the Ordnance Survey maps will necessarily require multi-disciplinary expertise.

The following set of comments contains information which is also provided in corresponding parameters below:

```
+Perthshire+
                                %LC: Perthshire, unspecified
+unspecified+

=DrummondRJean=                 %FN: DrummondRJean6130317
```

+Perthshire+ permits the grouping together of all documents localized in the county of Perthshire. Since this particular letter does not contain information about the place in which the letter has been written, this is pointed out by +unspecified+; in numerous letters there is a place-name here (e.g., +Perth+).

=DrummondRJean= allows the identification of all letters by a particular writer, exactly the same form of the name also occurring as the first part in the filenames of those letters (see the description of %FN: below).

%ST: stands for 'status' and describes whether a letter was transcribed in situ or whether the transcription is based on a photocopy or photograph of the original manuscript in the CSC archive (the acronym CSC occurs in one of the values of this parameter instead of ScotsCorr as a reflection of the earlier stages of the corpus project). Letters transcribed in situ have only been rechecked once.

%DA: specifies the date of a letter, in the order year (e.g. 1686), month (e.g. April) and day (e.g. 13). Note. Elsewhere in the list of parameters, zeroes or question-marks may appear as part of dates. If a particular piece of information is doubtful, a question mark may follow (686?0413 in the case of a doubtful year, 68604?13 in the case of a doubtful month, and 6860413? in the case of a doubtful day). Zeros are used when information is missing (6860400). In the case of undated letters, the first or second half of a century have been given as the approximate date, drawing on information about the birth and death of the writer (6000000 referring to the first half of the seventeenth century and 6500000 to the second). When a particular informant has written several letters at a particular date, Roman numerals have been added (e.g. 6860413I, 6860413II, etc.).

%CO: refers to 'contents'. In this file-initial parameter, information is restricted to naming the writer and the addressee and providing his or her title or profession at the moment of writing (e.g., Jean Drummond, Countess of Roxburgh, and Anna Livingston, Countess of Eglinton). No further information is provided, now that there are numerous online sources for studying the informants. Frequently, the online catalogues of the archives also contain summaries of the contents of the letters, especially that of the NRS.

%BI: for ‘biographical data’ focuses on providing references to earlier editions of the document. The parameter value ‘information unavailable’ points out that the compiler is not aware of the existence of an earlier edition of a particular letter (on differences between the original manuscript and nineteenth-century editions, for example, see Meurman-Solin 2013a).

%IF: stands for ‘informant, female’, the person who signed the letter. The other values of this parameter are **%IM:** ‘informant, male’ and **%IR:** ‘informant, royal’.

There are a few letters in the ScotsCorr in which there are two signatures. In these, the name of the person in whose hand the letter is written is positioned first in the description of the informant. It should be noted that in non-autograph letters the informant signing the letter is not its writer. The user is advised to use the parameters %IF/IM/IR: and %HD1/2: in conjunction, in order to distinguish informants represented by autograph letters from those whose own hand only appears in the signature (and sometimes also the letter-closing formula). For more information, see %HD: below.

The ScotsCorr informants have been described by extracting information from various sources, focusing on basic facts that can be directly related to the definition of the informants with reference to the variables of time, space, and social milieu. As stated in Section 3 Dimensions of space, time, and social milieu, time and space have been considered more important than other variables which have been viewed as relevant in recent research in historical sociolinguistics. Lack of balance as regards social stratification in the ScotsCorr, there being too few informants representing the lower social classes, has prevented the formalization of parameter values related to social milieu. In other words, being a linguist, the compiler has been reluctant to translate prosopographical information into a compartmentalized – and compartmentalizing – system of social indices without first consulting researchers in the fields of social and economic history and cultural studies. The provision of information without the suggestion that this information can be used in a straightforward way to ‘explain’ the linguistic findings is a very conscious policy in the ScotsCorr (cf. Meurman-Solin 2001a).

%AF: is an abbreviation of ‘addressee, female’. The other values of this parameter are **%AM:** ‘addressee, male’ and **%AR:** ‘addressee, royal’.

This information is usually based on the address written on one side of a folded letter. When the manuscript does not have an address, a suggestion is sometimes recorded in the entry in the catalogues. Since any address on the manuscript is transcribed at the end of a given text file as part of the letter, the user will know which source of information has been used. If the address itself has not been checked against manuscript, it is put in parentheses in the transcript.

%HD: provides information about hand-writing, also stating whether a letter is autograph or non-autograph. When there are autograph and non-autograph passages in a particular letter, two parameters are used: **%HD1:** for autograph and **%HD2:** for non-autograph. The pairs of comments {hand1>} … {<hand 1} and {hand2>} … {<hand 2} in the text indicate where the autograph and non-autograph passages begin and end. The numbers 1 and 2 do not refer to the order in which the two hands occur in a text; instead, hand 1 is always autograph and hand 2 non-autograph.

For example, in letters in which the body of the letter has been written by an amanuensis and the letter-closing formula and signature, for instance, by the informant, the text-initial parameters include the following information:

%HD1: autograph, letter-closing formula and signature
%HD2: non-autograph, secretary

%LC: for 'locality' specifies the county or district to which the writer has been localized and quotes the name of the place in which a particular letter was written. The following counties or districts occur as values of the parameter %LC:

Aberdeenshire

Angus

Argyllshire

Ayrshire

Borders (i.e. Border counties)

Fife

Invernessshire

Lanarkshire

Lothian

Moray

Perthshire

Ross (i.e. Cromarty and Ross)

South West (Dumfries and Galloway)

Stirlingshire

Sutherland

While the above-listed areas have been used as parameter values in the file-initial lists of language-external variables (as in +Perthshire+ and %LC: Perthshire in the example above), the Korp application also provides attributes for grouping these areas into much larger regions according to the following schema:

NORTH	Moray
	Invernessshire
	Sutherland
	Ross
NORTH-EAST	Aberdeenshire
	Angus

CENTRAL	Perthshire
	Lanarkshire
SOUTH-EAST	Fife
	Lothian
	Stirlingshire
	Borders
SOUTH-WEST	Argyllshire
	Ayrshire
	South-West

The writer's geographical origin is not indicated in the following three categories:

UNLOCALISED

COURT

PROFESSIONAL

Altogether 43 informants of the ScotsCorr have not been localized (c. 15,000 words), these being put in the category "Unlocalised". The parameter value "unlocalised" is used when the writer is only known by name, and further information about him/her remains as yet unmined. Instead of suggesting a region, two parameter values of %LC specify the writer's adherence to the royal court (%LC: Court) or a professional coalition (%LC: Professional). In the present version of the ScotsCorr, the latter category is heterogeneous, containing members of the clergy, the army and the legal profession, for example. The decision not to localize professional people is based on the general assumption that the language of members of the clergy, for example, will reflect a geographically defined variety only partly, if at all, being influenced by the shared properties of conventionalized professional discourse. Since the present version also has too few informants in the category "Professional" (c. 28,000 words), a more refined categorization is inappropriate. The five different larger regions are not evenly represented, the proportion of south-eastern informants being largest of all the geographically defined letter-writers:

North	25 per cent
North-East	14 per cent
Central	12.5 per cent
South East	35.5 per cent
South-West	13 per cent

In %LC the name of the region is followed by the name of the place in which a particular letter was written. This place-name has either been replaced by its modern equivalent or it is quoted in the form in which it appears in the original manuscript. The latter practice has been adopted only if a place-name remains unidentifiable, or the variant used by the letter-writer is ambiguous. Some variants have remained opaque to the ScotsCorp compiler and the letter-writer's variant of a particular place-name is given as a parameter value. Since some letters have been written on the

Continent, there are also foreign place-names, some of these also appearing in the list of parameter values in the form in which they occur in the letters. When names of castles, palaces, residences, or institutions appear in a letter (e.g., Stirling Castle, Holyrood House, Whitehall), these have been replaced by names of cities (Stirling, Edinburgh, London).

It is obvious that the regional distribution closely reflects the practical issues in the selection process (see Section 4.1 Selection of data for the ScotsCorr). The areas in the South-East are more densely demarcated, whereas in the North and the South-West localization is suggested with reference to larger areas. The fact that Sutherland is named reflects the compiler's interest in the Gordon family of Sutherland, and can also be explained by the easy access to the Sutherland deposits in the National Library of Scotland.

%FN: stands for 'filename' and thus gives the filename of the letter in the corpus. This filename functions as a reference to a particular letter, indicating where a particular occurrence listed in a concordance, for example, has been attested. The filename also functions as a reference attached to each occurrence in the Korp concordance. The filenames have been selected with the following practices as guidelines: earls are referred to in terms of their position in the line of succession, e.g., 10Angus, 11Angus, 12Angus for the 10th, 11th, and 12th Earls of Angus; the title Lord is represented by L in filenames (1LDuffus, the 1st Lord Duffus), the title Marquis by M (3MMontrose, 3rd Marquis of Montrose), Duke by D (1DGordon, the 1st Duke of Gordon), and Viscount by V (2VMontgomery, the 2nd Viscount of Montgomery). The information given in other filenames is structured in the following order:

CrichtonEElizJNabeth6740108 Family name, initial of a family a female informant is married to: **Crichton** married to the family of the Earls of **E**glinton

First name: CrichtonE**Elizabeth**6740108

Year of writing (a three-digit number): CrichtonEElizabeth**674**0108
Month of writing (a two-digit number): CrichtonEElizabeth674**01**08
Day of writing (a two-digit number): CrichtonEElizabeth6740**108**

When a particular informant has written several letters at a particular date, Roman numerals have been added (e.g., 6860413I, 6860413II, etc.).

%WC: is an abbreviation of 'word count'. The totals have been calculated by software counting space-separated tokens but excluding all punctuation marks and the editor's comments.

The comment {CSC 2007} points out that a tagged version of the letter was produced for the 2007 CSC, which is not available internationally.

The following table briefly summarises the properties of file-initial variables and their parameter values:

# 111	the identification number; non-continuous; arbitrary in the sense that it only reflects the

	order in which a particular text was submitted to tagging
%MS:	the catalogue number either at the National Records of Scotland (NRS) or the National Library of Scotland (NLS); in some rare cases the value may be “information unavailable”
543	the year 1543 in which the letter was written
0 0 0	a variable providing an option for adding later coordinates on the map of Scotland, which specify the geographical origin of the writer or the place in which the letter has been written
+Aberdeenshire+	the area the writer orginates from; if it has not been possible to specify the writer’s geographical origin, the value is “unlocalised”
+Huntly+	the place in which the letter has been written; if no place is mentioned in the letter, the value is “unspecified”
=KeithHElizabeth=	the informant’s name (her maiden name, the initial of her married name, her first name), in the form in which it also appears in the filename
%ST:	related to the circumstances in which the transcript was produced; “a copy in the CSC archive” refers to the fact that the transcript has been produced in Helsinki by using a xeroxcopy or photograph of the original manuscript, now deposited in the ScotsCorr archive; “transcribed in situ” refers to the transcript having been produced at the NRS or NLS using the manuscript original itself
%DA:	the detailed date of the manuscript (yyyy, name of the month, dd); sometimes part of this information is missing; if the letter is undated, an approximate date is given (1600 for 1600-1649, 1650 for 1650-1699, and 1700 for 1700-1749)
%CO:	the writer of the letter (by) and its addressee (to)
%BI:	information about previous editions of the letter; the value “information unavailable” states that the compiler is not aware of the existence of any previous editions
%IF:	the informant and the gender/rank (F for female, M for male, and R for royal); the name in the form in which it appears in the first part (by) of %CO:
%AR:	the addressee and the gender/rank (R for royal, M for male, F for female); the name in

	the form in which it appears in the second part (to) of %CO:
%HD1:	the first part states whether the letter is "autograph" or "non-autograph"; the second part specifies whether the script type is "secretary", "non-secretary" or "italic"; if only the signature, the initial term of address and/or the letter-closing formula are autograph, the second part lists these features; %HD2: is added as a variable to letters in which there are two different hands, the values being "non-autograph" followed by one of the above-mentioned script type categories
%LC:	the region in which the writer has been located (the same value as in +Aberdeenshire+ in # 111 below) followed by the place of writing (the same as in +Huntly+ in # 111 below); if the writer's origin is unknown, the value is "unlocalised" and if the place of writing is unknown, the value is "unspecified"
%FN:	the file name, which contains the name between = = (KeithHElizabeth, as in # 111 below), a three-digit year, the number of the month and the day of writing
%WC:	word count
{CSC 2007}	indicates that the letter has been tagged to be included in the 2007 version of the corpus (at present unavailable)

Example # 111 illustrates both the file-initial variables and the letter itself, followed by explanations of particular transcription practices applied to the letter and comments considered relevant:

111
%MS: NRS SP2/1:17
543
0 0 0
+Aberdeenshire+
+Huntly+
=KeithHElizabeth=
%ST: a copy in the CSC archive
%DA: 1543 August 16
%CO: by Elizabeth Keith, Countess of Huntly, to Mary of Lorraine, Queen Dowager
%BI: previously edited by Annie I. Cameron in the Correspondence of Mary of Lorraine, 17:20-21
%IF: Elizabeth Keith, Countess of Huntly
%AR: Mary of Lorraine, Queen Dowager
%HD1: autograph, secretary

%LC: Aberdeenshire, Huntly
%FN: KeithHElizabeth5430816
%WC: 225
{CSC 2007}

{outdented to the left>} Madame” {<<M> enlarged} I co*m%mend” my hartly Seruice to zour grace It pleis zour grace I haue resauit zour \ grace writing fra zour s*er%vand yis berar makand” me*n%tyoun” y=t= my lord gou*er%no*ur% hes rasit ane” \ cursing on~ my lord and” done `be his awyn” avyce to stop y=t= he cum” no=t= to zour grace at yis \ tyme” as zo*ur% grace Is Informyt Madame” I assuyr zour grace ze {cancellation} will fynd y=t= Informa*tio%ne” \ als*s% fals*s% as vy*er%is quhilk*is% hes bene” maid to zour grace abefoir yair Is na syk l*et%res cu*m%min \ on” my lord as zit as ze wrayt bot my lord” wes Informyt y=t= syk l*et%res wes to cum” and” \ hes gottin” ane” absolu*tio%ne” fra my lord cardinall In ave*n%tuyr of ye samy*n% Madame” beleif \ na vy*er% thing bot my lord” wil_ _be ye samy*n% man~ he p*ro%mist to zour grace And” hes gottin” \ greyt Laubo*uris% be ye gou*er%no*uris% waye to brak hy*m% fra zour purpois*s% and had ya bene” \ any alteratioun” of purposis I suld” no=t= haf falit till adu*er%teis zour grace w=t= deligens \ and” aye sal_ _be redy to do {del} syk {del} zour grace syk plesuyr and” s*er%uice as I ma at \ all tymes as knawis*s% god quha mot haue zour grace In keping et*er%nalye \ at huntlie ye xvj daye of august be {space} {a wide space vertically} \\ zour*is% grace h*u%m*i%ll Seruatrice {a wide space vertically} \\ Elizab=t= countes of Hwntlye {address>} To ye quenis grace {end}

The commentary in letter # 111:

{outdented to the left>} The beginning of the first line has been positioned in the left margin.

{<<M> enlarged} The initial character M is considerably larger than other capital letters in the text.

{cancellation} Unidentifiable characters or a word or words have been cancelled.

{del} … {del} An identifiable word has been cancelled.

{space} There is a space between words which is wider than the distance between words in general in this particular text.

{a wide space vertically} There is a space between chunks of text which may be indicative of paragraph structure or put apart from one another different discourse types (e.g., the body of the letter from letter-closing formulae or the letter itself from additional material appropriately reflecting contemporary politeness strategies)

{address>} the address of the letter usually on the reverse side or a separate sheet

{end} indicates the end of the contents of a particular file

The symbols in letter #111:

between the symbols *…% an emic representation of a contracted form is provided

a word-final ” indicates that there is a loop or flourish attached to the final character of a word

sal_ _be indicates that the two words are written together in the original

~ indicates that the word contains a contraction but what has been contracted remains ambiguous, i.e., allows an interpretation in more than one way

=t= indicates that there is a superscript t

For a detailed description of symbols and comments in the ScotsCorr, see Section 7.

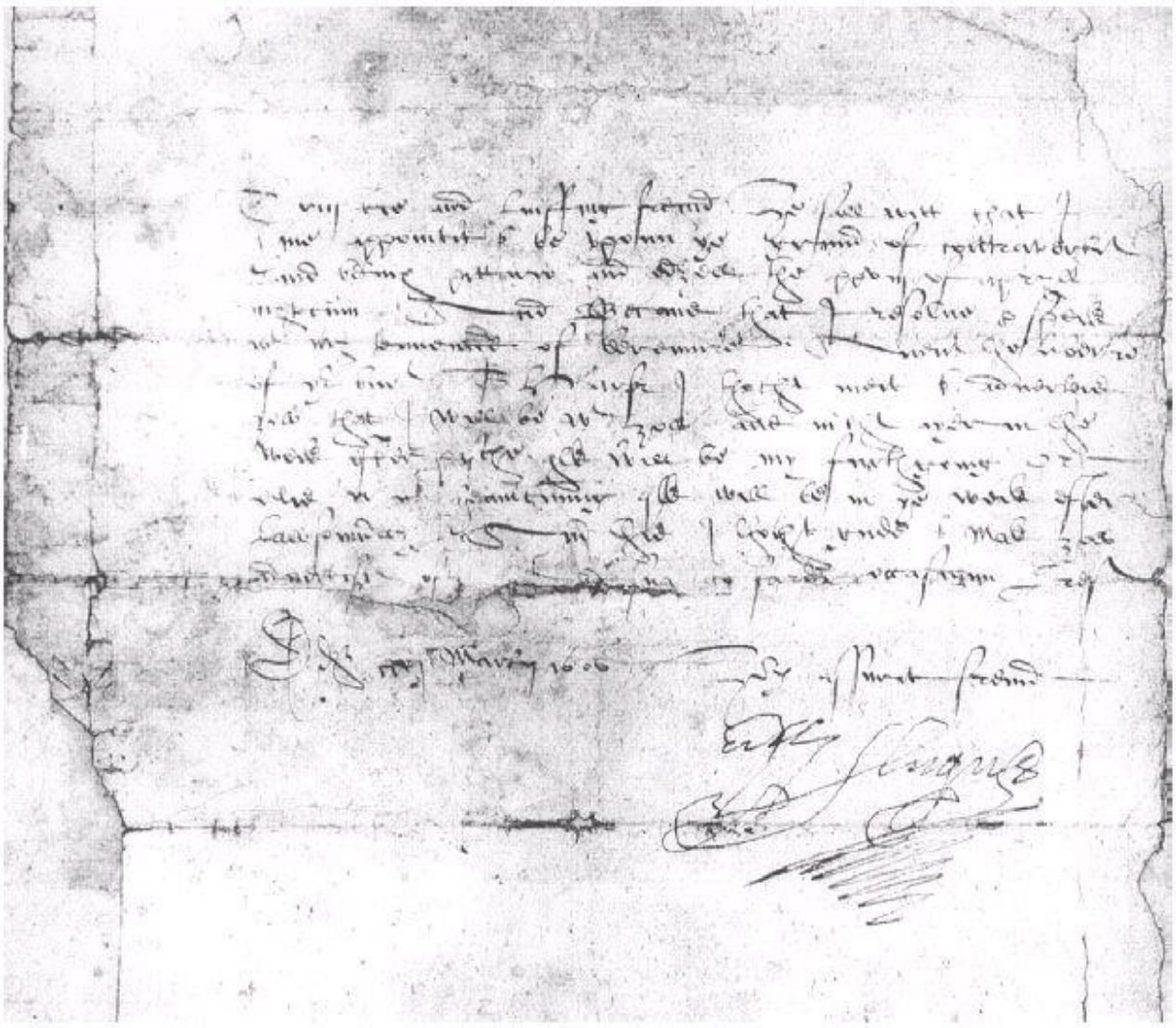
6 Visual prosody

The manuscript letters have been diplomatically transcribed and digitized, with detailed comments on non-linguistic features such as properties of the manuscript layout, paragraph structure, punctuation, particular character shapes, and spacing. In referring to these features I use the concept of visual prosody. The rationale in commenting on features of visual prosody is that these provide important information for linguistic analysis. In my own work on connectives (Meurman-Solin 2007b, 2011, 2012), for instance, the identification of sentence and discourse structure is often made possible through the examination of spacing and particular character shapes. An ideal solution would be to make digital images of the manuscripts available on the web, as well, so that the users of the ScotsCorr could check themselves what kind of additional information is provided by the visual prosody (see the illustrations in Meurman-Solin 2013a and b). The non-linguistic features commented on are as follows:

- physical condition (e.g., torn margin or damage by damp)
- number of folio
- line-break
- position of text (in margin, before or after the body of the letter)
- change of hand
- script type
- idiosyncratic features of a particular hand
- insertion; cancellation; correction
- punctuation
- spacing
- marked character shape
- paragraph structure

These features are discussed in Section 4.2 Transcription and digitization and in Meurman-Solin 2013a in particular, so in this section I will focus on illustrating marked character shapes and spacing in some examples and how these may affect linguistic analysis.

The digitized manuscript below illustrates what the historical letters in the ScotsCorr corpus look like.



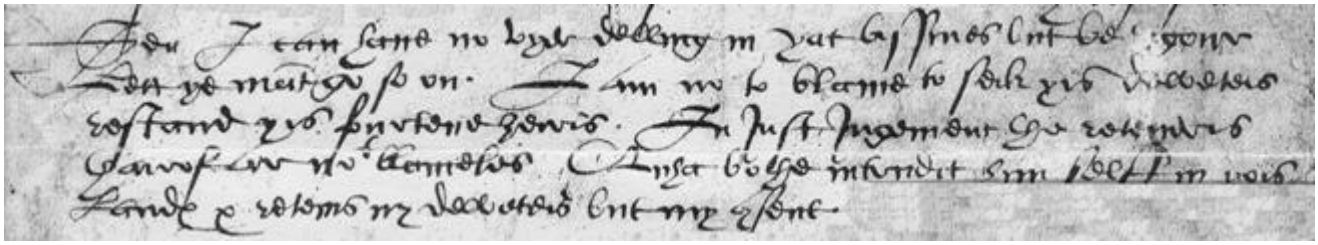
William Douglas, 10th Earl of Angus to Sir John Ogilvy of Inverquharity, Edinburgh, 31st March 1606. National Records of Scotland GD205/1/34. Published by the kind permission of the Trustees of Sir David Ogilvy of Inverquharity, Bt. (National Records of Scotland, GD205).

Trustie and luiffing freind Ze sall witt that I \ haue appointit to be vpoun ye grund of contravertit \ land betuix Pittarro and edzell the xxviij of apryll \ nixtocum **And** Becaus that I resolue to speik \ w=t= my tennentis of keremure Anent the libertie \ of y*ar% bur=t= **THairfor** I thocht meit to aduerteis \ zow that I will be w=t= zow ane nicht ayer in the \ weik efter pasche qlk will be my furthgoing Or - \ ellis in my heamcum*m%ing qlk will be in ye weik efter \ lawsounday **And** this I thocht gude to mak zow \ aduertisit of **And** sua to farder occasioun I rest {a space vertically} \\ Zo*uris% assurit freind \\ V D Erll Anguss {adjacent>} Ed*inburgh% xxxj Martij 1606 {end}

Some features of the visual prosody in this manuscript affect grammatical categorization, and certain variants also suggest categorial fuzziness. There are a number of instances of the text-structuring connective *and*, the function of which can be identified by the considerably extended shape of the initial character, which is clearly different from the shapes this character has elsewhere in the text. Notice also the larger size of both <T> and <H> in THairfor.

The following extract illustrates how evidence of the function of the relative pronoun WHO as a sentence-level reference signal can be chiefly inferred from visual prosody, being here used just like

a noun phrase or a demonstrative or personal pronoun. The sentence-level function is signalled by the space and the initial capital (notice also the spacing and capitalization in the sequence of sentences of the same length):



An extract from a letter by William Douglas, Marquis of Angus, 1642. National Records of Scotland GD205/1/34. Published by the kind permission of the Trustees of Sir David Ogilvy of Inverquharity, Bt. (National Records of Scotland, GD205).

\ Sen I can haue no vyer delling in yat bissines but be ri?gour \ Lett ye mat*er% go so on . {space} I am no to blame to seik yis deweteis \ restand yis fourtene zeiris . {space} In Iust Iugement the reteneris \ thairof ar no=t= blameles {space} **Quha** bothe intrudit him selff in yois \ Land*is% & retenis my deweteis but my *con%sent . (11Angus6420408)

In this example, visual prosody, spacing, and initial capitals prompt us to analyse the variation between the *Quha* shown in bold in the transcript and other ways of referring to the human participants of the text. In other words, spacing can be interpreted as suggesting that the passage consists of four separate sentences. *Quha* in this non-restrictive relative clause (notice the capitalization) can be replaced by a noun phrase, *these (above-mentioned) persons* or *men*. The relative is separated from the non-adjacently positioned NP (*the reteneris \ thairof*) by the predicate.

Let us examine all the human referents in the first third of this letter:

{hand 2>} Ry=t= assurit freind {space} Ze sall ressaue yir V*ar%ningis of Myne \ and direct" zour officiaris To vse yame lauchfullie in dew tyme aganis \ sic as appertenis . {space} As to fentrie I send zow ane decreit of removing \ obtenit be me aganis vmq=le= dauid dowglas : To p*er%sew fentrie in zo*ur% court \ as succeding in ye vice & violent occupatioun of dauid dowglas ffor the \ byrunis awand me be dauid dowglas & thois aganis quhome I haue \ obtenit yat decreit / {space} Thair is no Law nor ressone yat can purge \ any p*er%sonis fact or wrangous deid Quha maliciouslie intrudis him \ in possessioun of quhatsumewer p*er%sonis landis be fraud yat can debar \ ye herito*ur% or titular of thois landis frome his rentis & dewetes \ Bot ye p*er%sone intrusar of all necessitie is Lyabill to ye heretour \ for ye byrun maill*is% of ye Landis he wrangouslie intrudit him - \ selff in But *con%sent of ye heretour : fentrie is s*er%eff him selff \ of angus I dout no=t= he is Ignorant of yis forme of p*ro%ces \ And this I wilbe ans*er%abill to zow is ye costome & practik \ of yis cuntrie bothe befoir ye Lordis of sessioun & als in all vyeris \ Iudicatoreis w=t=hin yis kingdome quhairin I am certanlie {reff cancelled} \ resoluit IT is no freindlie delling to fentrie to intrude him in my \ Landis & bothe abstract" ye deweteis awand me be my taxisman And the \ dewetie sen his occupatioun . qlk is xiij or xiiij zeiris . {space} Thairfoir \ do me iustice & lett yir byrunis be adiwgit & decernit to me in zo*ur% \ court of all zeiris bypast to yis hour . {space} And I sall tak the {<2 words partly torn} \ most discreitt way in tyme {ins} cu*m%ing {ins} to sattill yat mater . with ressone . \ Sen I can haue no vyer delling in yat bissines but be ri?gour \ Lett ye mat*er% go so on . {space} I am no to blame to seik yis deweteis \ restand yis fourtene zeiris . {space} In Iust Iugement the reteneris \ thairof ar no=t= blameles {space} Quha

bothe intrudit him selff in yois \ Land*is% & retenis my deweteis but my *con%sent . giff pitmowis zour \ discharge for his byrun few maillis qlk salbe sufficient for him \ & he sall haue my discharge? *con%forme to zouris quhen he pleiss*s% \ I houp ze salbe ane rewlar of all my pepill in thois boundis \ of kerimure & sie y=t= no wrang be tane nor done be any of ya*m% \ to vyeris . And ressaue omocheis few maill*is% quhome I haue wrettin \ to giff zow yame & giff him zour discharge y*ar%of sen the dait \ of his chairto*ur% I rest and sall remane {space} \\ Zour assurit freind {<hand 2} \\ {hand 1>} W=m= Douglas {<hand 1} {adjacent>} {hand 2>} Ed*inburgh% 8 {italic>} aprile {<italic} \ 1642 {<hand 2} \ {ins} {hand 1>} what siluer yow gait giue to Condie \ and wreitt to me what yow giue him \ and if ye Lerd off Fntrie do a dewtie go \ fordwar w=t= him accordingly as I haue \ wreittin . {ins} {<hand 1} {end}

A careful reading of the letter shows that these retainers have been named and repeatedly referred to in the context preceding the extract. In fact, they have been discussed in great detail, both with reference to wrong-doing they are personally responsible for and to what the law says about any person who infringes on another person's rights. Thus, the non-restrictive relative clause in the extract contains given information. The passage suggests a reading: 'Alas! In just judgement, the retainers thereof [of these lands] are not blameless. After all, they did intrude in those lands and retained my duties without my consent.' (In the preceding context, the reference signals alternate with *who* with generic reference 'whoever' and 'whatever persons', which may explain the use of *himself* here, i.e., this use is a reflection of the pattern of co-reference in *whoever*, *any person who* and *he*.)

Another example will illustrate that, in addition to spacing and the use of capitals, the interpretation as a sentence-level reference signal is suggested by the distance between the antecedent, or as Huddleston and Pullum (2002) suggest, the anchor, and the relative element. The following extract contains a relative structure as a time relator:

ffor samekle as **ye tent day off** {torn except for initial <f>; Fraser has: Febru>} **f???\ar nixttocu*m%** is assignit to me & all vyeris To produc~ befoir \ his ma=tie= and counsall o*ur% clames & titillis quhair we acclame \ honour*is% and places in parliame*n%t*is% and g*ene%rall counsallis . Or \ neuir to be hard y*ar%eftir : {space} **Qlk day** godwilling I purpois \ to keip (1601 William Douglas, 10th Earl of Angus; 10Angus6010112)

The space and the colon in the manuscript before *Qlk day* and the initial capital of the pronoun *Qlk* 'which' suggest that the relative structure may function as a sentence-level signal of anaphoric reference. The antecedent realised by *ye tent day off f???ar nixttocum* in this example is at quite a distance from *Qlk day*. The non-adjacent positioning of the relative element and its anchor is quite frequent in historical data: the position of the relative WHICH with inanimate reference is non-adjacent in 27 per cent of the occurrences in the CSC 2007 data (for further information, see Meurman-Solin 2007b).

Given that punctuation in historical documents is not sufficiently regularized to allow the reconstruction of clause and sentence structure, these illustrations permit us to draw the conclusion that a thorough understanding of implicit and explicit connectivity, as well as visual prosody, is necessary for us to learn to identify these structures. (Rydén 1966: xliii points out that in his data, relatives representing the *relativische Anknüpfung* type are 'often preceded by a full stop or other marks of heavy punctuation'. Nevertheless, he concludes that the inconsistencies attested in punctuation prevent us from considering these relative links as a discrete category of their own.) However, the assessment of the relevance of these visual features is by no means easy in individual

cases, and converging evidence of various kinds will have to be provided to create valid criteria for the analysis of clause and sentence structure in data of the present kind. This procedure resembles the analysis of spoken language (from sound rather than transcripts), in which prosodic features play a significant role. The analysis of manuscript letters written by relatively inexperienced as well as more competent and stylistically highly professional writers requires similar tools to that of recorded spoken data. Developing such tools and methods is one of the tasks of the research project this database is a product of.

7 Symbols and comments in the ScotsCorr

In the compiler/transcriber's practice of annotating the texts, both symbols and comments have been used. Symbols are either used in the same way as punctuation marks, with space on both sides of the symbol, or integrated in the words, with no cooccurrence of adjacent signs such as brackets allowed. In contrast, comments are always put in curly brackets, both within words and when used as independent comments, either related to a word or words by means of an arrow (< or >) or without any link with the immediate context. For a detailed discussion of symbols and comments, see Sections 7.1 and 7.2. For consulting the practices of using symbols and comments in alphabetical order, see the auxiliary databank Symbols and Comments in the ScotsCorr.

7.1 Symbols in the ScotsCorr Transcripts

Please notice that some of these symbols may also occur in the original manuscript in a different position from that described by the transcriber below. A comment about this has been added to the descriptions (see – and ~, for instance).

7.1.1 *Within words*

- at the end of a line either to mark an empty space or to signal text structure, i.e., the end of a chunk of discourse and the beginning of the next. Note: A hyphen in the manuscripts may also be used in the same functions as in modern English, in compounds and to divide words not only at line-ends but, often repeatedly, at the beginning of lines (e.g., pay-\-ment).

~ at the end of a word indicates that the word contains a contraction but what has been contracted remains ambiguous

” at the end of a word indicates that there is a loop or flourish attached to the final character of a word which may be interpreted either as a stylistic feature of the script type or as suggesting a variant word-form

#	in word fragments which do not allow their reading as words; the symbol indicates where the missing part is (e.g., al#, where the two initial characters could be the beginning of any number of different words)
*…%	indicate that the characters between them are contracted in the original manuscript, the contraction being usually marked by a flourish of some kind, a particular character shape (in *per% and *con%, for example), or a line above or at the end of the contracted word; in the case of s*s% to represent the character ß in the manuscript, there is only a single character between this pair of symbols (expens*s%)
=…=	indicates that the character or characters between the symbols is in superscript in the original manuscript (e.g., kny^{t} 'knight' is digitized as kny=t=)
=	occurs in the manuscripts, varying with a hyphen, in the function of dividing a word between two lines, not only at line-ends but, often repeatedly, at the beginning of lines (e.g., pay=\=ment)
?	indicates that the immediately preceding character is unclear, irregular, or ambiguous; there is usually a comment with an arrow following the word which suggests why the character does not allow straightforward interpretation
??	indicates that in the position in which the two question-marks occur a character remains illegible
???	indicates that in the position in which the three question-marks occur two or more characters remain illegible

7.1.2 *Independent symbols*

\	a backward slash indicates line-break
\\	two backward slashes indicate that a new paragraph follows
-	at the end of a line either to mark an empty space or to signal text structure, i.e., the end of a chunk of discourse and the beginning of the next. Note: A hyphen in the manuscripts may also be used in the same functions as in modern English, in compounds and to divide words not only at line-ends but, often repeatedly, at the beginning of lines (e.g., pay-\-ment).
~	at the end of a line either to mark an empty space or to signal text structure, i.e. the end of a chunk of discourse and the beginning of the next
/	the virgule, with varying shapes resembling a forward slash, used as a punctuation mark especially in the pre-1650 letters

7.1.3 *Splitting two combined words*

…_ _... separate two words written together in the manuscript (e.g., salbe sal_ _be)

7.2 Comments in the ScotsCorr transcripts

By position, comments in curly brackets can be grouped into two categories, those within words and those positioned between words in the text. By presence or absence of an explicit link, they can also be classified into two types: those that indicate by either '<' or '>' whether the comment is on an element in the immediately preceding word or words or in the immediately following context. Below a third set of criteria is applied to describing the comments used in the ScotsCorr. Section 7.2.1 discusses the function of comments depicting the physical or visually detectable features of the manuscript original. Section 7.2.2 examines the often interrelated language-external and linguistic features of handwriting or script type which call for a comment for the user to assess the validity of a particular occurrence.

7.2.1 Comments on the physical or visually detectable features of the manuscript original

Language-external comments on the manuscript original provide the user with information about the physical condition of the manuscript and the text, focusing on features which may have negatively affected the transcription process. In other words, in the case of a damaged manuscript, comments such as {torn} or {<blurred} are inserted into the text to make the user aware that a detail in the original text is partly or fully unrecoverable or the reading of a word or passage in very pale ink is doubtful. A comment without '<' refers to something which is damaged or torn completely; since emendation is not allowed, the torn part will remain as an unfilled gap in the running text, the comment {torn} indicating its position. A comment with '<' indicates that the reading of the preceding item is doubtful due to the physical feature specified by the comment. Thus, the comment {<torn} usually occurs in contexts in which the preceding item contains question-marks, indicating that some characters are partly torn in the manuscript original (e.g., dou?n?e {<partly torn} marks the incomplete shapes of <u> and <n> in the manuscript). The function of these comments is to permit the user to assess the quality of the data and to be aware of the lower validity of these unclear attestations.

7.2.2 Comments on script type and handwriting

These comments mark any idiosyncratic features of handwriting that may lead to ambiguity. Since ambiguous characters in a particular hand make the identification of a given linguistic variant doubtful, the comments may contain both language-external (i.e., script-specific) and linguistic (i.e., related to potential patterns of variation) information. With particularly untidy or badly-formed hands, it may sometimes be impossible to decipher a word, and the doubtful characters in the reading provided have been signalled with an immediately following question-mark and a comment, which may suggest an alternative reading.

Even in regular hands, <t> and <c> often have similar shapes, and comments are used to inform the user about the alternative reading (see Section 4.3.3). In hands that do not properly distinguish between <a> and <o>, the linguistic item and the comment have the following structure: ha?me {<or <o>}. If this ambiguity applies to the majority of shapes a particular pair of characters have in a particular hand, a comment about this may also be positioned before the body of the text in the file (e.g,. {shapes of <a> and <o> are sometimes indistinguishable}). Since some degree of ambiguity may be caused by writing <i>, consistently or inconsistently, without a dot or <t> without a

horizontal stroke, comments such as {<with …} or {<without…} have been introduced (e.g., t?ake {<without a horizontal stroke}). Similarly, when a particular character has a feature which is otherwise absent in contemporary script, for example a dot above <I> in the first-person subject pronoun, this is pointed out by adding the comment {<with a dot}.

In addition to unclear realizations of individual characters, the idiosyncratic ductus in a hand may be reflected in a compressed realization of particular sequences of characters; for example, a sequence of minims may be represented by a wavy line. In these cases, the approximate realizations of individual characters have been signalled as such with question-marks, and the comment {<compressed} or {<reduced} follows. The comment {<compressed} refers to a sequence of characters which, assessed by both shape and size, are smaller than the realisations of the same characters in the letter or elsewhere in the same writer's hand, whereas the comment {<reduced} refers to a sequence of characters whose shape is reduced to a curve or even a straight line; these reduced shapes are especially frequent in inflectional morphemes, such as *-ing* in the progressive, at line-ends, and lexical morphemes, such as variants of *-tion*/*-sion*.

Especially in terms of address at the beginning of a letter or in letter-closing formulae and at the beginning of new paragraphs or chunks of discourse enlarged characters appear in the manuscripts. In the attached comments in the transcripts the enlarged feature or features have been specified. The comment {<<…> enlarged} indicates that the size, often also the shape, of a character is clearly larger than that of the same character in upper case elsewhere in the text (e.g., My {<<M> enlarged) Dear Lord). The comment {<<…> extended} indicates that the shape, often also the size, of a character is clearly extended, or stretched out to cover more space, in comparison with the same lower- or upper-case character elsewhere in the text (e.g., and {<<a> extended} at the beginning of a new chunk of discourse).

For information about change of hand, see Section 5 Language external information in the text-files.

7.2.3 Comments on layout

The number of folios a letter consists of varies considerably, but, in this manual, it is not possible to provide a detailed statistical account of this variation. A letter may begin on the right side of a sheet and continue on the left, the reverse side, or a new sheet or sheets. It may also cover the first sheet, continue in its margin and end there, or continue from the margin of the first sheet to the second (for examples, see Meurman-Solin 2013 a and b. The total number of sheets can be counted by using the comments {f1}, {f2}, etc, and the layout by the comments {f1r}, {f1v}, {f2r}, etc. Because of problems related to lack of information as regards how the Xerox-copies of the manuscripts have been produced by the archives, comments such as {f1v} and {f2v} may refer either to the left or the reverse side.

Use of space or indentation may highlight the structure of a letter by signaling that a particular chunk of text has a particular discourse function. Another function is the marking of paragraph structure or the beginning of a chunk of text which has a discourse function different from the preceding one (Meurman-Solin 2012; see also Section 6). Since the development of how paragraphs are marked in letters is a highly interesting topic, the comment {space} is used in the transcripts to indicate that there is a (horizontal) space in the text which is clearly wider than that between words. The comment {left indenture>} allows to find all the occurrences in which the text is indented from the left margin, while {outdented to the left>} points out that the beginning of the line has been positioned in the left margin, i.e., outside the position of the body of the letter.

When a letter continues in margin, the comment {in margin>} is used; this comment is frequently followed by {direction changes>} because, due to restrictions of space, the direction of writing changes in margin as compared with the preceding text.

Comments such as {centred>}, {a space vertically}, and {a wide space vertically} are particularly frequent in the function of providing information about the position of a term of address at the beginning of a letter and that of letter-closing formulae. The comment {address>} allows the identification of who a letter is written to, the position of this text being usually on the front (visually the best) side of an often numerous times folded sheet.

7.2.4 Comments on cancellation, correction, deletion, and insertion

Since epistolary prose in manuscript letters is by nature unedited, any cancellation or correction remains visible, whether a strikethrough is used or a correction is imposed on an erroneously written item. When what has been cancelled can no longer be read, the independent comment {cancellation} is used, wheras an unclear correction which remains illegible is commented on by {an unclear correction}. When a legible word or words have been cancelled, the item or items are put between {del} … {del} 'deleted'. This pair of comments can also appear within words, a single {del} being used if the cancelled element appears word-finally. The comments {<corrected} and {<an unclear correction} may follow a word in which a question-mark or question-marks indicate that the word remains ambiguous, and the comment provides the reason for this ambiguity. Insertions are marked by the pair {ins} … {ins}; a single {ins} is used if the inserted element appears word-finally.

{=…} = and a modern English equivalent follow a word which has been abbreviated by only using its initial character (e.g., L {=lordship}, M {=majesty})

{address>} precedes the address written on the front side of the folded letter

{blurred} a word or words which are blurred because of damage by damp, for example, or pale ink, and therefore illegible,

{<blurred} a word or words which, because of damage by damp, for example, or pale ink are partly blurred but still legible; a character which remains ambiguous is followed by a question-mark; a character which remains illegible is replaced by two question-marks, whereas a sequence of presumably more than one character (judged by space available) is replaced by three question-marks

{cancellation} indicates that part of a word or words have been cancelled, so that because of being incomplete or because of thick strikethroughs or unclear correction they can no longer be read; this is the case with false starts consisting of a character or two, or in cases where a character or characters have been crossed out as a correction, for instance

{centred>} the text is positioned in the middle of a line or indented from the left and right, or, as is often the case with letter-closing formulae and the signature particularly in post-1600 letters, aligned to fit the right margin

{<compressed}

refers to a sequence of characters which, assessed by both shape and size, are smaller than the realisations of the same characters in the letter or elsewhere in the same writer's hand

{damaged} a word or words which are damaged, so that they remain illegible

{<damaged}
a word or words which are damaged but still partly legible; a character which remains ambiguous is followed by a question-mark; a character which remains illegible is replaced by two question-marks, whereas a sequence of presumably more than one character (judged by space available) is replaced by three question-marks

{del} … {del}
a deleted word or words; when a correction follows a deleted item or items, this is usually provided by inserting the correction; in the transcript this results in the sequence: {del} I {del} {ins} we {ins}

{direction changes>}
the direction of writing changes as compared with the preceding text

{<<…> enlarged}
the size, often also the shape, of a character is clearly larger than that of the same character in upper case elsewhere in the text (e.g., My {<<M> enlarged) Dear Lord)

{<<…> extended}
the shape, often also the size, of a character is clearly extended, or stretched out to cover more space, in comparison with the same lower- or upper-case character elsewhere in the text (e.g., and {<<a> extended} at the beginning of a new chunk of discourse)

{f1r} folio 1 the right side or the first page
{f1v} folio 1 the left side or the reverse side

{hand1>} … {<hand 1} indicate the beginning and the end of autograph text in a letter written in two differenct hands

{hand2>} … {<hand 2} indicate the beginning and the end of non-autograph text in a letter written in two differenct hands

{in margin>} the text continues in margin

{ins} … {ins} an inserted word or words; when a correction follows a deleted item or items, this is usually provided by inserting the correction; in the transcript this results in the sequence: {del} I {del} {ins} we {ins}

{ins}…(ins} {<in margin} there is an insertion (e.g., a word or words) in margin

{left indenture>} the text is indented from the left margin

{<or <…>}	suggests an alternative reading where a particular character is ambiguous (e.g., sa?me {<or <o>})
{outdented to the left>}	the beginning of the first line has been positioned in the left margin
{<reduced}	refers to a sequence of characters whose shape is reduced to a curve or even a straight line; these reduced shapes are especially frequent in inflectional morphemes, such as *-ing* in the progressive, at line-ends, and lexical morphemes, such as variants of *-tion*/*-sion*
{space}	there is a (horizontal) space in the text which is clearly wider than that between words
{a space vertically}	there is a vertical space between lines which is clearly one line wider than that between lines elsewhere in the letter
{torn}	a word or words which are damaged, so that they remain illegible
{<torn}	a word or words which are partly torn but still, at least partly, legible; a character which remains ambiguous is followed by a question-mark; a character which remains illegible is replaced by two question-marks, whereas a sequence of presumably more than one character (judged by space available) is replaced by three question-marks
{a wide space vertically}	there is a vertical space between lines which is clearly wider than two lines elsewhere in the letter
{<with …}	a particular character has a feature which is otherwise absent in contemporary script (e.g., a dot above <I> in the first-person subject pronoun is pointed out by adding the comment {<with a dot}
{<without a horizontal stroke}	<t> is written without a horizontal stroke, the comment pointing out that, in the case of that particular word, this practice may make the reading of this particular character ambiguous

References

Aitken, A.J. 1971. Variation and variety in written Middle Scots. In: *Edinburgh Studies in English and Scots*, edited by A.J. Aitken, Angus McIntosh & Hermann Pálsson, 177-209. London: Longman.

Anderson, Jean, David Beavan & Christian J. Kay. 2007. SCOTS: Scottish Corpus of Texts and Speech. In: *Creating and Digitizing Language Corpora. Vol. 1: Synchronic Databases*, edited by Joan Beal, Karen Corrigan & Hermann Moisl, 17-34. Basingstoke: Palgrave Macmillan.

Barton, David & Nigel Hall (eds.). 1999. *Letter-Writing as Social Practice* (Studies in Written Language and Literacy, 9). Amsterdam & Philadelphia: Benjamins.

Benson, Phil. 2001. *Ethnocentrism and the English Dictionary*. London and New York: Routledge.
Bergs, Alexander. 2005. *Social Networks and Historical Sociolinguistics. Studies in Morphosyntactic Variation in the Paston Letters (1421–1503)*. Berlin and New York: Mouton de Gruyter.
Dareau, Marace. 2004. DOST: A significant instance of historical lexicography. In: *New Perspectives on English Historical Linguistics*, edited by Christian J. Kay, Carole Hough & Irené Wotherspoon, 49–64. Amsterdam: Benjamins.
Dareau, Marace. 2005. The history and development of DOST. In: *Perspectives on the Older Scottish Tongue. A Celebration of* DOST, edited by Christian J. Kay & Margaret A. Mackay, 18-37. Edinburgh: Edinburgh University Press.
Daybell, James. 2001. *Early Modern Women's Letter-Writing in England, 1450–1700*. Basingstoke: Palgrave Macmillan.
Daybell, James. 2012. *The Material Letter in Early Modern England: Manuscript Letters and the Culture and Practices of Letter-Writing, 1512–1635*. Basingstoke: Palgrave Macmillan.
Devitt, Amy J. 1989. *Standardizing Written English. Diffusion in the Case of Scotland 1520-1659*. Cambridge: Cambridge University Press.
Dossena, Marina. 2004. Towards a corpus of nineteenth-century Scottish correspondence. *Linguistica e Filologia* 18: 195-214.
Dossena, Marina & Gabriella Del Lungo Camiciotti (eds.) 2012. *Letter Writing in Late Modern Europe*. Amsterdam: Benjamins.
Dossena, Marina. 2012. The Study of Correspondence: Theoretical and Methodological Issues. In: *Letter Writing in Late Modern Europe*, edited by Marina Dossena & Gabriella Del Lungo Camiciotti, 13-30. Amsterdam: Benjamins.
Dossena, Marina. 2013. Ego Documents in Scottish Corpora: The Contribution of Nineteenth-century Letters and Diaries to the Study of Language History. In: *Language in Scotland: Corpus-based Studies*, edited by Wendy Anderson, 91-111. Amsterdam: Rodopi.
Fitzmaurice, Susan. 2002. *The Familiar Letter in Early Modern English*. Amsterdam & Philadelphia: Benjamins.
Fleischman, Suzanne. 2000. Methodologies and Ideologies in Historical Linguistics: On Working with Older Languages. In: *Textual Parameters in Older Languages*, edited by Susan C. Herring, Pieter van Reenen & Lene Schøsler, 33-58. Amsterdam: Benjamins.
Herring, Susan C., Pieter van Reenen & Lene Schøsler (eds.) 2000. *Textual Parameters in Older Languages*. Amsterdam: Benjamins.
Hickey, Raymond. 2000. Processing corpora with *Corpus Presenter*. *ICAME Journal* 24: 65-84.
Hickey, Raymond. 2003. *Corpus Presenter: Software for Language Analysis*. Amsterdam: Benjamins.
Houston, R. A. 1985. *Scottish Literacy and the Scottish Identity. Illiteracy and society in Scotland and northern England, 1600-1800*. Cambridge: Cambridge University Press.
Huddleston, Rodney & Geoffrey K. Pullum. 2002. *The Cambridge Grammar of the English Language*. Cambridge: Cambridge University Press.
Johnston, Paul 1997. Older Scots Phonology and its Regional Variation. In: *The Edinburgh History of the Scots Language*, edited by Charles Jones, 47-111. Edinburgh: Edinburgh University Press.
Kay, Christian J. & Margaret A. Mackay (eds.) 2005. *Perspectives on the Older Scottish Tongue. A Celebration of* DOST. Edinburgh: Edinburgh University Press.
Kytö, Merja (comp.). 1996 [1991]. *Manual to the Diachronic Part of the Helsinki Corpus of English Texts: Coding Conventions and Lists of Source Texts*. Third ed. Helsinki: Department of English, University of Helsinki.
Laing, Margaret. 2004. Multidimensionality: Time, Space and Stratigraphy in Historical Dialectology. In: *Methods and Data in English Historical Dialectology*, edited by Marina

Dossena & Roger Lass, 49-96 (Linguistic Insights. Studies in Language and Communication, 16). Bern: Lang.
Laing, Margaret & Keith Williamson. 2004. The Archaeology of Medieval Texts. In: *Categorization in the History of English*, edited by Christian J. Kay & Jeremy J. Smith, 85-145. Amsterdam: Benjamins.
LALME = *A Linguistic Atlas of Late Mediaeval English* (1986), 4 vols, edited by Angus McIntosh, Michael L. Samuels & Michael Benskin, with the assistance of Margaret Laing and Keith Williamson. Aberdeen: Aberdeen University Press.
eLALME = Benskin, Michael, Margaret Laing, Vasilis Karaiskos & Keith Williamson. An Electronic Version of A Linguistic Atlas of Late Mediaeval English. http://www.lel.ed.ac.uk/ihd/elalme/elalme.html
Lass, Roger. 2004. Ut custodiant litteras: Editions, corpora and witnesshood. In: *Methods and Data in English Historical Dialectology*, edited by Marina Dossena & Roger Lass, 21-48 (Studies in Language and Communication 16). Bern: Lang.
Lehmann, Christian. 1988. Towards a typology of clause linkage. In: *Clause Combining in Grammar and Discourse*, edited by John Haiman & Sandra A. Thompson, 181-225. Amsterdam & Philadelphia: Benjamins.
Marshall, Rosalind K. 1983. *Virgins and Viragos*. *A History of Women in Scotland 1080 to 1980*. London: Collins.
Meurman-Solin, Anneli. 1993a. *Variation and change in early Scottish prose*. *Studies based on the Helsinki Corpus of Older Scots*. (Annales Academiae Scientiarum Fennicae, Diss. Humanarum Litterarum, 65). Helsinki.
Meurman-Solin, Anneli. 1993b. Periphrastic and auxiliary *do* in early Scottish prose genres. In: *Early English in the computer age: explorations through the Helsinki Corpus*, edited by Matti Rissanen, Merja Kytö & Minna Palander-Collin, 235-251 (Topics in English Linguistics, 11). Berlin: Mouton de Gruyter.
Meurman-Solin, Anneli. 1995. A New Tool: The Helsinki Corpus of Older Scots (1450-1700). *ICAME Journal* 19: 49-62.
Meurman-Solin, Anneli. 1999. Letters as a Source of Data for Reconstructing Early Spoken Scots. In: *Writing in Nonstandard English*, edited by Irma Taavitsainen, Gunnel Melchers & Päivi Pahta, 305-322. Amsterdam: Benjamins.
Meurman-Solin, Anneli. 2000a. Change from above or from below? Mapping the *loci* of linguistic change in the history of Scottish English. In: *The Development of Standard English, 1300-1800: theories, descriptions, conflicts*, edited by Laura Wright, 155-170. Cambridge: Cambridge University Press.
Meurman-Solin, Anneli. 2000b. On the conditioning of geographical and social distance in language variation and change in Renaissance Scots. In: *The History of English in a Social Context. A Contribution to Historical Sociolinguistics*, edited by Dieter Kastovsky & Arthur Mettinger, 227-255. Berlin: Mouton de Gruyter.
Meurman-Solin, Anneli. 2000c. Geographical, socio-spatial and systemic distance in the spread of the relative *who* in Scots. In: *Generative Theory and Corpus Studies: A Dialogue from 10ICEHL*, edited by Ricardo Bermúdez-Otero, David Denison, Richard M. Hogg & C. B. McCully, 417-438. Berlin: Mouton de Gruyter.
Meurman-Solin, Anneli. 2001a. Structured Text Corpora in the Study of Language Variation and Change. *Literary and Linguistic Computing* 16/1: 5-27.
Meurman-Solin, Anneli. 2001b. Women as Informants in the Reconstruction of Geographically and Socioculturally Conditioned Language Variation and Change in the 16th and 17th Century Scots. *Scottish Language* 20: 20-46.

Meurman-Solin, Anneli. 2002. The progressive in early Scots. In: *English Historical Syntax and Morphology. Selected Papers from 11ICEHL*, edited by Teresa Fanego, María José López-Couso & Javier Pérez-Guerra, 203-229. Amsterdam: Benjamins.

Meurman-Solin, Anneli. 2003. Corpus-based Study of Older Scots Grammar and Lexis. In: *The Edinburgh Companion to Scots*, edited by John Corbett, J. Derrick McClure & Jane Stuart-Smith, 170-196. Edinburgh: Edinburgh University Press.

Meurman-Solin, Anneli. 2004a. Towards a Variationist Typology of Clausal Connectives. Methodological Considerations Based on the Corpus of Scottish Correspondence. In: *Methods and Data in English Historical Dialectology*, edited by Marina Dossena & Roger Lass, 171-197 (Linguistic Insights. Studies in Language and Communication, 16). Bern: Lang.

Meurman-Solin, Anneli. 2004b. From Inventory to Typology in English Historical Dialectology. In: *New Perspectives on English Historical Linguistics*, Volume I: *Syntax and Morphology*, edited by Christian J. Kay, Simon Horobin & Jeremy Smith, 125-151. Amsterdam: Benjamins.

Meurman-Solin, Anneli 2004c. Data and Methods in Scottish Historical Linguistics. In: *The History of English and the Dynamics of Power*, edited by Ermanno Barisone, Maria Luisa Maggioni & Paola Tornaghi, 25-42. Alessandria: Edizioni dell'Orso.

Meurman-Solin, Anneli 2005. Women's Scots: Gender-Based Variation in Renaissance Letters. In: *Older Scots Literature*, edited by Sally Mapstone, 424-440. Edinburgh: John Donald.

Meurman-Solin, Anneli. 2007a. Annotating variational space over time. In: *Annotating variation and change*), edited by Anneli Meurman-Solin & Arja Nurmi (Studies in Variation, Contacts and Change in English, 1). Helsinki: VARIENG. http://www.helsinki.fi/varieng/journal/volumes/01/meurman-solin/

Meurman-Solin, Anneli. 2007b. Relatives as sentence-level connectives. In: *Connectives in the History of English*, edited by Ursula Lenker & Anneli Meurman-Solin, 255-287 (Current Issues in Linguistic Theory, 283). Amsterdam & Philadelpia: Benjamins.

Meurman-Solin, Anneli. 2011. Utterance-initial connective elements in early Scottish epistolary prose. In: *Connectives in Synchrony and Diachrony in European Languages*, edited by Anneli Meurman-Solin & Ursula Lenker (Studies in Variation, Contacts and Change in English, 8). Helsinki: VARIENG. http://www.helsinki.fi/varieng/journal/volumes/08/meurman-solin/

Meurman-Solin, Anneli. 2012. Early Modern English Dialects. In: *Historical English Linguistics: An International Handbook*, vol. 1, edited by Alexander Bergs & Laurel Brinton, 668-684 (Handbücher zur Sprach- und Kommunikationswissenschaft/ Handbooks of Linguistics and Communication Science 34.1-34.2.). Berlin and New York: De Gruyter Mouton.

Meurman-Solin, Anneli. 2013a. Visual prosody in manuscript letters in the study of syntax and discourse. In: Principles and Practices for the Digital Editing and Annotation of Diachronic Data, edited by Anneli Meurman-Solin & Jukka Tyrkkö (Studies in Variation, Contacts and Change in English, 14). Helsinki: VARIENG. http://www.helsinki.fi/varieng/journal/volumes/14/meurman-solin_a/

Meurman-Solin, Anneli. 2013b. Features of layout in sixteenth- and seventeenth-century Scottish letters. In: *Principles and Practices for the Digital Editing and Annotation of Diachronic Data*, edited by Anneli Meurman-Solin & Jukka Tyrkkö (Studies in Variation, Contacts and Change in English, 14). Helsinki: VARIENG. http://www.helsinki.fi/varieng/journal/volumes/14/meurman-solin_b/

Meurman-Solin, Anneli 2013c. Taxonomisation of features of visual prosody. In: *Principles and Practices for the Digital Editing and Annotation of Diachronic Data*, edited by Anneli Meurman-Solin & Jukka Tyrkkö (Studies in Variation, Contacts and Change in English, 14). Helsinki: VARIENG.

http://www.helsinki.fi/varieng/journal/volumes/14/meurman-solin_c/
Meurman-Solin, Anneli & Arja Nurmi. 2004. Circumstantial Adverbials and Stylistic Literacy in the Evolution of Epistolary Discourse. In: *Language Variation in Europe. Papers from ICLaVE* 2, edited by Britt-Louise Gunnarsson, Lena Bergström, Gerd Eklund, Staffan Fridell, Lise H. Hansen, Angela Karstadt, Bengt Nordberg, Eva Sundgren & Mats Thelander, 302-314. Uppsala: Universitetstryckeriet.
Meurman-Solin, Anneli & Päivi Pahta. 2006. Circumstantial adverbials in discourse: a synchronic and a diachronic perspective. In: *The Changing Face of Corpus Linguistics*, edited by Antoinette Renouf & Andrew Kehoe, 117-141 (Proceedings of the 24 th International Conference of the International Computer Archive of Modern and Medieval English). Amsterdam & Atlanta, GA: Rodopi.
Meurman-Solin, Anneli & Jukka Tyrkkö. 2013. Introduction. In: *Principles and Practices for the Digital Editing and Annotation of Diachronic Data*), edited by Anneli Meurman-Solin & Jukka Tyrkkö (Studies in Variation, Contacts and Change in English, 14). Helsinki: VARIENG.
http://www.helsinki.fi/varieng/journal/volumes/14/introduction.html
Milroy, James. 1999. The consequences of standardization in descriptive linguistics. In: *Standard English. The Widening Debate*, edited by Tony Bex & Richard J. Watts, 16-39. London: Routledge.
Nevala, Minna. 2004. *Address in Early English Correspondence. Its Forms and Socio-Pragmatic Functions* (Mémoires de la Société Néophilologique de Helsinki, 64). Helsinki: Société Néophilologique.
Nevala, Minna & Arja Nurmi. 2013. The *Corpora of Early English Correspondence* (CEEC400). In: *Principles and Practices for the Digital Editing and Annotation of Diachronic Data*, edited by Anneli Meurman-Solin & Jukka Tyrkkö (Studies in Variation, Contacts and Change in English, 14). Helsinki: VARIENG.
http://www.helsinki.fi/varieng/journal/volumes/14/nevala_nurmi/
Nevalainen, Terttu & Helena Raumolin-Brunberg. 1996. The Corpus of Early English Correspondence. In: *Sociolinguistics and Language History*. *Studies Based on the Corpus of Early English Correspondence*, edited by Terttu Nevalainen & Helena Raumolin-Brunberg, 39-54. Amsterdam: Rodopi.
Nevalainen, Terttu & Helena Raumolin-Brunberg. 2003. *Historical Sociolinguistics*. London: Longman.
Nurmi, Arja. 2002. Does size matter? The *Corpus of Early English Correspondence* and its sampler. In: *Variation Past and Present. VARIENG Studies on English for Terttu Nevalainen*, edited by Helena Raumolin-Brunberg, Minna Nevala, Arja Nurmi & Matti Rissanen, 173-184 (Mémoires de la Société Néophilologique de Helsinki, 61). Helsinki: Société Néophilologique.
Palander-Collin, Minna. 1999. *Grammaticalization and Social Embedding: I THINK and METHINKS in Middle and Early Modern English* (Mémoires de la Société Néophilologique de Helsinki, 55). Helsinki: Société Néophilologique.
Palander-Collin, Minna & Minna Nevala (eds.). 2005. *Letters and Letter Writing*. *European Journal of English Studies (EJES)* 9/1.
Rissanen, Matti & Jukka Tyrkkö. 2013. The Helsinki Corpus of English Texts (HC). In: *Principles and Practices for the Digital Editing and Annotation of Diachronic Data*, edited by Anneli Meurman-Solin & Jukka Tyrkkö (Studies in Variation, Contacts and Change in English, 14). Helsinki: VARIENG. http://www.helsinki.fi/varieng/journal/volumes/14/rissanen_tyrkko/
Rydén, Mats. 1966. *Relative Constructions in Early Sixteenth Century English*. *With Special Reference to Sir Thomas Elyot* (Acta Universitatis Upsaliensis. Studia Anglistica Upsaliensia, 3). Stockholm: Almqvist & Wiksell.

Sairio, Anni & Minna Nevala. 2013. Social dimensions of layout in eighteenth-century letters and letter-writing manuals. In: *Principles and Practices for the Digital Editing and Annotation of Diachronic Data*, edited by Anneli Meurman-Solin & Jukka Tyrkkö (Studies in Variation, Contacts and Change in English, 14). Helsinki: VARIENG. http://www.helsinki.fi/varieng/journal/volumes/14/sairio_nevala/

Schneider, Gary. 2005. *Culture of Epistolarity: Vernacular Letters and Letter Writing in Early Modern England, 1500–1700*. Newark, DE: University of Delaware Press.

Taavitsainen, Irma & Päivi Pahta. 2013. The Corpus of Early English Medical Writing (1375–1800) – a register-specific diachronic corpus for studying the history of scientific writing. In: *Principles and Practices for the Digital Editing and Annotation of Diachronic Data*, edited by Anneli Meurman-Solin & Jukka Tyrkkö (Studies in Variation, Contacts and Change in English, 14). Helsinki: VARIENG. http://www.helsinki.fi/varieng/journal/volumes/14/taavitsainen_pahta/

Williamson, Keith. 1992/93. A Computer-aided Method for Making a Linguistic Atlas of Older Scots. *Scottish Language* 11-12: 138-173.

Williamson, Keith. 2000. Changing Spaces: Linguistic Relationships and the Dialect Continuum. In: *Placing Middle English in Context*, edited by Irma Taavitsainen, Terttu Nevalainen, Päivi Pahta & Matti Rissanen, 141–179. Berlin: Mouton de Gruyter.

Williamson, Keith. 2001. Spatio-Temporal Aspects of Older Scots Texts. In: *Scottish Language* 20: 1-19.

Williamson, Keith. 2004. On Chronicity and Space(s) in Historical Dialectology. In: *Methods and Data in English Historical Dialectology*, edited by Marina Dossena & Roger Lass, 97-136. Bern: Lang.

Williamson, Keith. 2005. DOST and LAOS: a Caledonian symbiosis?. In: *Perspectives on the Older Scottish Tongue*. *A Celebration of* DOST, edited by Christian J. Kay & Margaret A. Mackay, 179-198. Edinburgh: Edinburgh University Press.

Williamson, I. Keith. 2012. Historical Dialectology. In: *Historical English Linguistics: An International Handbook*, vol. 1, edited by Alexander Bergs & Laurel Brinton, 1421-1437 (Handbücher zur Sprach- und Kommunikationswissenschaft/ Handbooks of Linguistics and Communication Science 34.1-34.2.) . Berlin and New York: De Gruyter Mouton.

References to the Helsinki corpora of Scots:

HCOS = Anneli Meurman-Solin comp. (1995) *Helsinki Corpus of Older Scots*, 1450-1700. http://www.helsinki.fi/varieng/CoRD/corpora/HCOS/index.html

CSC = Anneli Meurman-Solin, comp. *Corpus of Scottish Correspondence (1542-1708)*. First edition (2003). http://www.helsinki.fi/varieng/CoRD/corpora/CSC/index.html

CSC= Anneli Meurman-Solin, comp. *Corpus of Scottish Correspondence (1542-1708)*. Second edition (2007). http://www.helsinki.fi/varieng/CoRD/corpora/CSC/index.html

ScotsCorr= Anneli Meurman-Solin, Research Unit for the Study of Variation, Contacts and Change in English (VARIENG), Department of Modern Languages, University of Helsinki: The Helsinki Corpus of Scottish Correspondence 1540–1750 (2017) [text corpus]. - FIN-CLARIN [referred to on dd.mm.yyyy]. Available in Kielipankki, the Language Bank of Finland, at http://urn.fi/urn:nbn:fi:lb-201411071