A New Tool: The Helsinki Corpus of Older Scots (1450–1700)

Anneli Meurman-Solin
University of Helsinki

1. Introduction to the study of Scots

In the diachronic study of the regional varieties of the English language, the study of the Scottish English variety has now reached a stage in which it greatly benefits from the computerization of a wide range of texts, and from the recent developments towards more sophisticated methods in the field of corpus linguistics. As early documents of regional varieties are usually less readily available than texts representing varieties that have reached the status of a standard, the international distribution of computerized material on regional varieties can be assumed to arouse more interest in these often unexplored territories. Corpora of this kind also allow scholars working in other fields to refer to converging or contradictory evidence outside their main area of interest without much extra effort.

There are two research projects under way on Older Scots: one, focusing at present on the fourteenth and fifteenth centuries, is at the Institute for Historical Dialectology, University of Edinburgh; and the other, at present covering the period 1450–1700, is in the Department of English, University of Helsinki. The person responsible for the project in Edinburgh is Dr. Keith Williamson; the present writer has compiled the Helsinki Scots Corpus, as a member of the Helsinki Corpus team, supervised by Professor Matti Rissanen[1]. The general aim in both projects is to provide computer-readable material for tracing linguistic variation and change over time. The texts in the Edinburgh corpus are based on manuscripts of entire texts, chiefly representing non-literary material, i.e. local records and legal documents. One of the principal objectives of the Edinburgh project is to make a linguistic atlas of Older Scots at different periods. The Helsinki Corpus uses extracts from early prints or later editions of the texts, and is structured by coding information about sociohistorically relevant extralinguistic variables. The aims of the Helsinki Corpus will be discussed below.

Variation analysis was introduced into Scottish studies as early as 1971, when Professor A.J. Aitken published his pioneering article on variation and variety in Middle Scots. To mention just a few others, Amy Devitt and Michael Montgomery have stressed the importance of including texts outside the basic canon in their research material, but the text corpora used by them have, to my knowledge, remained private. Except for a selection of texts computerized for the purposes of *The Dictionary of the Older Scottish Tongue*, and now available through the Oxford Text Archive, texts representing the early stages of the Scottish variety have not been within easy reach of international scholars.

2. The compilation of the Corpus

The Helsinki Corpus of Older Scots has been compiled as a supplement to the diachronic part of the Helsinki Corpus of English Texts: Diachronic and Dialectal (see Kytö 1993, Rissanen *et al*. 1993, Kytö *et al*. 1994:33–39, 53–63, 73–79). The Scottish texts were selected according to the same principles of sociohistorical variation analysis as the main corpus, and the computer format, parameter coding and editorial and typographical conventions are also the same. In general terms, this supplement provides material for studying the last stages of the differentiation of the northern English dialect, the rise of a distinctive Scottish variety of English and the anglicization process of Scots. During the planning process, the chief aim was to compile a corpus that would allow comparisons of different geographical varieties of the English language, particularly those varieties available in a computer-readable form in the Helsinki Corpus of English Texts, and in the supplement of early American English.

The Scots Corpus offers an opportunity to consult a much wider range of texts representing the variety than has previously been available to researchers of the history of Scots. It contains approximately 830,000 words of running text; the periodization and number of words per period are as follows:

1450–1500	85,100
1500–1570	201,800
1570–1640	305,900
1640–1700	241,400
Total	834,200

The texts represent fifteen different prose genres: acts of Parliament, burgh records, trial proceedings, histories, biographies, travelogues, diaries, pamphlets, educational treatises, scientific treatises, handbooks, private letters, official letters, sermons and the Bible (for further information, see the table in the Appendix). The focus on non-literary genres, including those which may be claimed to reflect usages of spoken language or, more appropriately, to favour stylistically marked variants that are typical of informal settings, will also make it possible to consider the influence of the social roles and social networks of the authors and the addressees.

3. Experimenting with the pilot version of the Corpus

Meurman-Solin (1993a) comprises an extensive introduction to the pilot version of the Scots Corpus and six studies based on the corpus material. The volume includes a detailed discussion of the principles of compilation and the extralinguistic features of the texts, and specifies the different parameter values of these variables. Five of the six studies that follow the introduction illustrate the application of different methods to a number of topics such as verb morphology, periphrastic *do*, and frequencies and distributions of Scottish and English variant spellings; one of these studies looks at co-occurrence patterns of linguistic features in Douglas Biber's (1988) Factor 1, 'Involved' as against 'Informational' production, and shows how the different genres in the Scots Corpus are positioned on this dimension.

The studies show for instance that regional variation *within* the area of present-day Scotland cannot be ignored in any descriptive work on the Scottish variety; that, in the sixteenth century, printing gave support to the establishment of a Scottish English norm, whereas, in the seventeenth century, it led to the preference of anglicized variants; that especially such language-external variables as the author's sex and the nature of the participant relationship in letters condition the choice between markedly Scottish and English variants. The chronology, pace and direction of change depend on the linguistic features analysed; there are important differences between developments such as the differentiation and anglicization processes related to spelling variation, the varying strength of the so-called subject-verb constraint, and the introduction and spread of periphrastic *do* in Scottish.

Another interesting dimension is the position of the Scottish representatives of medieval and Renaissance prose genres in the history of English literature. In a number of studies based on the Corpus (Meurman-Solin forthcoming a – c), the compiler has looked at viewpoint marking and degree of involvement as opposed to degree of informational density and the prominence of descriptive elements, and analysed the co-occurrence patterns of linguistic features assumed to be relevant in the identification of genre-specific characteristics (cf. Biber 1988, Biber and Finegan 1989, Chafe 1985, Simpson 1993). Important differences have been attested between the Scottish and the English representatives of genres, also as regards the preference for integrated or fragmented structures, so that for example the frequency of descriptive adjectives, particularly in modifier positions, is often twice as high in English as in Scottish texts.

4. Revising and expanding the pilot Corpus

Experience based on the intensive use of the pilot Corpus by its compiler has suggested certain changes in the version of the Helsinki Corpus of Older Scots prepared for international distribution. Besides administrative texts such as local records of Scottish burghs, genres containing material of chiefly local interest, for example histories, turned out to be linguistically conservative. To increase the representativeness of the Corpus in this respect, samples from one early-sixteenth-century history and *The Carnwath Barony Court Book* have been added; the computerized version of the latter was kindly made available by Dr. Williamson. As the difference between so-called Central or East Mid Scots, and Northern or North-East Scots was found to be significant (Meurman-Solin 1993a: 136–137), extracts from the burgh records of Aberdeen were included. The most important addition is the inclusion of a selection of Acts of the Parliaments of Scotland to represent each of the four time periods.

In addition to the expanded selection of laws and records, the pre-1500 texts now contain a couple of new shorter prose works, but this subperiod remains less representative. However, the focus of the Edinburgh project on the earlier stages of the Scottish variety will solve the problem of the scarcity of early prose texts in the Helsinki Corpus. In the period 1500–1570, a diary by John Lesley, the historian, allows the study of idiolectal variation in two texts representing different genres. Such comparisons are available also in the idiolect of Sir Patrick Waus, whose language can be studied in his private letters, written when he was a schoolboy to his mother, his official letters and a short travelogue. As early letters are very rare in the published correspondences of Scottish noble families, two new specimens dating from 1400 and 1405 have also been included. Moreover, a number of official letters from the first two decades of the sixteenth century were felt to complement in a useful way those in *The Correspondence of Mary of Lorraine*, which date from the period 1542– 1571.

New editions that have recently been made available replace the older ones used in the pilot version; a particularly important achievement is Jonathan Glenn's new edition of Gilbert Hay's *The Buke of Knychthede* and *The Buke of the Governaunce of Princis*. The Woodrow Society edition of Robert Bruce's sermon has been replaced by the more reliable version in the Laing edition. The dialogue passages, recording the trial against George Wishart, in Pitscottie's history (*a* 1578), and James VI's *Counterblaste to Tobacco* have not been omitted in the second version, although the former is based on a black-letter tract printed in London, and the latter was written after the Union of the Crowns and was therefore addressed also – or perhaps primarily – to a London audience. The special character of the intentionally scotticized New Testament should be acknowledged by the user of the Corpus.

5. Extralinguistic variables

As stated above, the parameter values given to the different extralinguistic variables have been discussed in Meurman-Solin (1993a), but a further analysis of language-external evidence related to the texts has suggested a number of changes to the pilot version. The texts are grouped into genres by extralinguistic criteria, and no direct correspondence between these genres and a classification into text types by linguistic criteria can be assumed. As regards text categories, the parameter value 'argumentative', which does not occur in the Helsinki Corpus of English Texts, is used for pamphlets in the Scots Corpus. As in the Helsinki Corpus, double labels have been used, when appropriate, so that for example the allegorical element in *The Complaynt of Scotland* and Buchanan's *Chamaelon* has been stressed by labelling them as 'argumentative/imaginative narration'. There is not always a direct correspondence between genres and categories; George Sinclair's *Satan's Invisible World Discovered*, which is a selection of stories about witches, has been labelled as education because of its instructive aim, but it represents the text category 'non-imaginative narration'.

Yet another variable, used exclusively in the Scottish supplement, provides information about whether a text was immediately printed or not. This was considered important, as printing has been assumed to be one of the most important accelerators of anglicization processes in

Scots (cf. however Meurman-Solin 1993a:137–148).

The variables ‘interaction’ and ‘setting’ have been shown to condition the clustering of specific sets of linguistic features in particular genres and text categories, but further study is required to interrelate the social and communicative functions of genres and the linguistic variation reflecting these functions. In contrast, the nature of the ‘participant relationship’ in letters, which has been described in terms of parameter values such as ‘intimate up/equal/down’, ‘distant up/equal/down’, clearly influences the choice of variants, so that for example a preference for conservative Scottish variants is typical of letters written to an addressee who is, from the writer’s perspective, in a socially inferior position.

The system of parameter values for the variable ‘audience’ aims at giving a more detailed description of the general function of the texts in their social and cultural context. Despite its tentativeness, this more detailed system was coded into the Scots Corpus, as the labels ‘professional’, ‘non-professional’ and ‘unspecified’, used in the main corpus, would have meant singling out only one text, a late-seventeenth-century scientific treatise, as a text written to a professional audience. While it has not yet been possible to show significant correlation between these variables and markedly Scottish or English variants, the system of parameter values turned out to be justified and useful in studies aiming at the reconstruction of a grammar of point of view (Meurman-Solin forthcoming a – c) or in those discussing degree of involvement as opposed to informational density.

Women’s language is represented by approximately 20,000 words of running text in the Scots Corpus; on the basis of this limited evidence, female writers tend to prefer conservatively Scottish variants, and a considerable time lag has been attested in the anglicization of their language.

6. Ongoing work

Because of the interesting findings presented in the pilot studies, the compiler of the present corpus has started selecting texts for a new corpus of texts by Scottish female writers. This corpus will cover the period 1500–1800 and include a considerably wider selection of letters, but also texts representing other genres, such as essays on various topics (for example educational treatises), travelogues, diaries, autobiographies, fiction and plays.

The first experiments with the grammatical tagging of the Scots Corpus have started in cooperation with Keith Williamson, who has developed the computer programs for the linguistic analysis of medieval vernacular texts, both early Middle English and Older Scots (for further information, see Laing in Kytö *et al*. 1994). He is also responsible for the first application of these programs to text material in the Helsinki format. The first text profiles have been successfully produced. Instead of preselected data, assumed to be diagnostic of the Scottish variety on the basis of earlier research, it is now possible to use larger quantities of evidence, which have been put in a manageable format by processing and listing them in various ways. It seems that the cooperation between the Edinburgh and Helsinki projects will make it possible to provide a tagged corpus of Scots ranging from the very earliest written documents of the variety to the early eighteenth century.

The new Corpus is now available from the Norwegian Computing Centre for the Humanities (see the order form accompanying this journal). Hopefully, the corpus will later be included in a CD-ROM disk containing other text corpora, together with the Helsinki Corpus of English Texts, as part of the "ICAME Collection of English Language Corpora" (for details on technical matters and distribution, see Kytö 1993). Despite the possible weaknesses that still remain in the present version, it was important to make the current version available without further delay. The compiler welcomes comments and suggestions on this current version to allow future assessment and revision.

Correspondence: Department of English
P.O. Box 4 (Hallituskatu 11)
FIN-00014 University of Helsinki
Finland

E-mail: meurmansolin@cc.helsinki.fi

Note

1. I would like to thank Professors Matti Rissanen, Merja Kytö and A.J. Aitken for their unfailing support in the compilation of the corpus, as well as all my colleagues in the Helsinki Corpus project for most inspiring discussions. I would like to acknowledge the very valuable assistance offered by Kirsi Heikkonen and Arja Nurmi, who in their unenviable position as research assistants always patiently and generously helped me complete the present version of the corpus. I have also greatly benefited from the recently launched cooperation with Dr. Keith Williamson.

References

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

Biber, Douglas. 1988. *Variation across speech and writing*. Cambridge, New York, etc.: Cambridge University Press.

Biber, Douglas and Edward Finegan. 1989. Styles of stance in English: Lexical and grammatical marking of evidentiality and affect, *TEXT*, 9–1:93–124.

Chafe, Wallace L. 1985. Linguistic differences produced by differences between speaking and writing, in *Literacy, language and learning: The nature and consequences of reading and writing*, ed. David R. Olson, Nancy Torrance and Angela Hilyard, 105–123. Cambridge: Cambridge University Press.

Devitt, Amy J. 1989. *Standardizing written English. Diffusion in the case of Scotland, 1520–1659*. Cambridge, New York, etc.: Cambridge University Press.

Kytö, Merja. 1993. *Manual to the diachronic part of the Helsinki Corpus of English Texts: Coding conventions and lists of source texts.* 2nd edition. Helsinki: Helsinki University Press.

Kytö, Merja, Matti Rissanen and Susan Wright (eds.). 1994. *Corpora across the centuries. Proceedings of the First International Colloquium on English Diachronic Corpora*. Amsterdam and Atlanta, GA: Rodopi.

Laing, Margaret. 1994. The linguistic analysis of medieval vernacular texts: Two projects at Edinburgh, in Kytö, Merja, Matti Rissanen and Susan Wright, pp. 121–141.

Meurman-Solin Anneli. 1993a. *Variation and change in early Scottish prose. Studies based on the Helsinki Corpus of Older Scots*. Annales Academiae Scientiarum Fennicae, 65. Helsinki.

Meurman-Solin Anneli. 1993b. Introduction to the Helsinki Scots Corpus. In Matti Rissanen, Merja Kytö and Minna Palander-Collin (eds.), 75–82.

Meurman-Solin Anneli. 1994. The Helsinki Corpus of Older Scots. In Kytö, Merja, Matti Rissanen and Susan Wright (eds.), 53–63.

Meurman-Solin Anneli. Forthcoming a. Towards reconstructing a grammar of point of view. Textual roles of adjectives and open-class adverbs in Early Modern English Texts. In Rissanen, Matti, Merja Kytö and Kirsi Heikkonen (eds.).

Meurman-Solin Anneli. Forthcoming b. Marking of stance in Early Modern English imaginative narration. A paper read at the conference on Narrative Strategies in Early English Fiction. Salzburg, October 1994.

Meurman-Solin Anneli. Forthcoming c. Point of view in Scottish and English genre styles. A paper read at the 8th International Conference of English Historical Linguistics. Edinburgh, September 1994.

Meurman-Solin Anneli. Forthcoming d. On differentiation and standardization processes in early Scots. In Jones, Charles (ed.), *History of the Scots Language*.

Meurman-Solin Anneli. Forthcoming e. Text profiles and dialect maps in the reconstruction of Renaissance Scots.

Montgomery, Michael. Forthcoming. The evolution of verb concord in Scots.

Rissanen, Matti. 1989. Three problems connected with the use of diachronic corpora, *ICAME Journal*, 13:16–19.

Rissanen, Matti, Merja Kytö and Minna Palander-Collin (eds.). 1993. *Early English in the computer age. Explorations through the Helsinki Corpus*. Berlin and New York: Mouton de Gruyter.

Rissanen, Matti, Merja Kytö and Kirsi Heikkonen (eds.). Forthcoming. *English in transition. Diachronic corpus studies in variation*.

Simpson, Paul. 1993. *Language, ideology and point of view*. London and New York: Routledge.

Appendix

Texts in the revised and expanded version of the Helsinki Scots Corpus arranged by prototypical text category and genre, with word counts (please note that there may be slight discrepancies between these counts and those given in the version available for international distribution). X = parameter value unspecified.

PROTOTYPICAL TEXT CATEGORY	TEXT	WORD COUNT
GENRE		
SC0 1450–1500		
STATUTORY		
LAW	*Acts of Parliament* (1455-1483)	27,300
STATUTORY		
RECORD	*Aberdeen Burgh Records* (1450-1489)	4,900
SECULAR INSTRUCTION		
EDUCATION	G. Hay, *Prose MS* (1456)	28,800
	Porteous of Noblenes (1490)	3,800
RELIGIOUS INSTRUCTION		
EDUCATION	*Craft of Deyng* (1450)	3,200
	Vertewis of the Mess (*c* 1460)	800
	Dicta Salomonis (1460?)	6,400
	Irland, *Meroure of Wyssdome* (1490)	8,300
X		
CORRESPONDENCE: OFFICIAL	Various writers (1400-1405)	1,500
SC1 1500–1570		
STATUTORY		
LAW	*Acts of Parliament* (1525-1555)	30,400
STATUTORY		
RECORD	Stirling Burgh Records (1519–1529)	10,800
	Carnwath Barony Court (1523-1542)	2,500
	Edinburgh Burgh Records (1540–1551)	3,100
	Peebles Burgh Records (1555–1573)	19,700
	Aberdeen Burgh Records (1519–1556)	10,300

PROTOTYPICAL TEXT CATEGORY	TEXT	WORD COUNT
GENRE		
STATUTORY		
TRIAL	*Sheriff Court Book of Fife* (1515–1522)	8,700
RELIGIOUS INSTRUCTION		
EDUCATION	Gau, *The Richt Vay to Heuine* (1533)	9,500
NON-IMAGINATIVE NARRATION		
HISTORY	*Mar Lodge Boece* (*c* 1533)	21,800
ARGUMENTATIVE		
PAMPHLET	Lamb, *Resonyng* (1550)	6,600
	Q. Kennedy, *Eucharistic Tracts* (1561)	14,600
ARGUMENTATIVE/IMAGINATIVE NARRATION		
PAMPHLET	*Complaynt of Scotland* (1549–1550)	8,700
X		
TRIAL	*St.Andrews Kirk Sessions* (1559–1561)	5 ,500
	Criminal Trials (1561–1567)	7,900
X		
CORRESPONDENCE: PRIVATE	Patrick Waus (1540)	1,300
X		
CORRESPONDENCE: OFFICIAL	*Corr. of Mary of Lorraine* (1542–1560)	33,000
	Various writers (1515-1521)	3,000
X		
BIBLE	Nisbet's scotticized *New Testament* (1520)	4,400
SC2 1570–1640		
STATUTORY		
LAW	*Acts of Parliament* (1587-1621)	49,700
STATUTORY		
RECORD	Stirling Burgh Records (1600–1608)	10,100
	Aberdeen Burgh Records (1590–1620)	10,000

PROTOTYPICAL TEXT CATEGORY	TEXT	WORD COUNT
GENRE		
EXPOSITORY HANDBOOK	Huntar, *Weights and Measures* (1624)	6,900
SCIENCE	Skeyne, *Descriptioun of the Pest* (1568) Skeyne, *Descriptioun of the Well* (1580)	9,100
SECULAR INSTRUCTION		
EDUCATION	James VI, *Basilicon Doron* (1598)	19,900
RELIGIOUS INSTRUCTION		
SERMON	Fergusson (1571)	6,700
	R. Bruce (1590–1591)	6,200
	J. Row (1638)	2.500
NON-IMAGINATIVE NARRATION		
HISTORY	Lesley (1570)	10,600
	Pitscottie (*a* 1578)	10,800
	Moysie (1590–1598)	10,600
TRAVELOGUE	Lithgow, *Trauayles* (1632)	12,400
	Patrick Waus (1587)	2,000
DIARY	Lesley (1571)	8,300
	Melville (of Kilrenny) (1600–1610)	15,700
	Birrel (1605)	12,700
	Johnston (1632–1660)	9,300
AUTOBIOGRAPHY	Melville (of Halhill) (1610)	8,600

PROTOTYPICAL TEXT CATEGORY	TEXT	WORD COUNT
GENRE		
ARGUMENTATIVE		
PAMPHLET	Fowler, *Answer to Hamiltoun* (1590)	6,100
	James VI, *Counterblaste to Tobacco* (1604)	5,000
	Birnie, *Blame of Kirk-buriall* (1606)	5,800
ARGUMENTATIVE/IMAGINATIVE		
NARRATION PAMPHLET	Buchanan, *Chamaelon* (1570)	4,600
ARGUMENTATIVE		
TRIAL	*Wishart Trial* (1562-63/a 1578)	3,800
X		
TRIAL	*St.Andrews Kirk Sessions* (1589–1592)	17,500
	Criminal Trials (1576–1591)	10,000
	Trial David Roy (1601)	3,900
X		
CORRESPONDENCE: PRIVATE	Various writers (1569–1635)	15,300
X		
CORRESPONDENCE: OFFICIAL	Various writers (1587–1631)	11,800
SC3 1640-1700		
STATUTORY		
LAW	*Acts of Parliament* (1661-1686)	40,800
STATUTORY		
RECORD	*Stirling Burgh Records* (1667–1680)	10,000
	Aberdeen Burgh Records (16 –16)	10,000
EXPOSITORY		
HANDBOOK	Skene, *Of Husbandrie* (1669)	1,700
	Reid, *Scots Gard'ner* (1683)	10,900
SCIENCE/OTHER	Sinclair, *Hydrostaticks* (1672)	15,200
	Sinclair, *Natural Philosophy* (1683)	

PROTOTYPICAL TEXT CATEGORY		TEXT	WORD COUNT
	GENRE		
RELIGIOUS INSTRUCTION			
	SERMON	Welsh, *Sermon John XX* (1679)	4,500
NON-IMAGINATIVE NARRATION			
	EDUCATION	Sinclair, *Satan's Invisible World* (1685)	7,800
	HISTORY	Spalding (c 1650)	15,600
	TRAVELOGUE	*Prince of Tartaria* (1661)	3,300
		Lauder, *Journals* (1665–1676)	10,000
	DIARY	Lamont (1649–1671)	7,000
		A. Brodie (1652–1680)	12,200
		J. Brodie (1680–1685)	6,600
		Andrew Hay (1659–1660)	10,000
		Cunningham (1673–1680)	9,200
	BIOGRAPHY	Somerville (1679)	6,900
		Turner (1632–1670)	7,800
AGUMENTATIVE			
	PAMPHLET	*Presbyterian Eloquence* (1692)	8,600
		Apology for the Clergy (1692–1693)	8,500
X			
	TRIAL	*Tryal Philip Standsfield* (1688)	8,500
X			
	CORRESPONDENCE: PRIVATE	Various writers (1659–1705)	18,600
X			
	CORRESPONDENCE: OFFICIAL	Various writers (1660–1708)	7,700