Studia Anglica Posnaniensia 47, 2-3, 2012
doi: 10.2478/v10121-012-0001-0

LINGUISTICS

LONG LEXICAL BUNDLES AND STANDARDISATION IN HISTORICAL LEGAL TEXTS[1]

JOANNA KOPACZYK

Adam Mickiewicz University, Poznań

ABSTRACT

Standardisation on the level of text is visible in the employment of stable and fixed expressions for a specific textual purpose. When gauging the extent of standardisation in texts, one of the parameters which should be taken into consideration is the length of such stable patterns. Since it is more difficult, and therefore rarer, to reproduce long chunks of text in an unchanged form, such a practice points towards greater standardisation. To explore the textual behaviour of long fixed strings in legal texts, this paper concentrates on long lexical bundles built out of eight consecutive elements (8-grams) and their frequency and function in historical legal texts. The database for this pilot paper comprises two collections of legal and administrative texts written in Scots between the fourteenth and the sixteenth century. The research results point to a considerable degree of textual standardisation throughout the corpus and to the most prominent functions of long repetitive chunks in historical legal discourse.

1. Lexical bundles and the study of historical discourse

A *lexical bundle* is a recurrent string of words in a text, regardless of its semantic or phrasal structure. Lexical bundles are extracted by a computer program which moves through the text one word-boundary at a time, searching for strings of a given length repeated in exactly the same form and arrangement. This method stems from the research into grammar teaching methods (Biber 1997) and was developed by Biber and his associates in their seminal grammar of English (Biber et al. 1999) and then in subsequent publications, dealing

[1] This research is supported by the Polish Ministry of Science and Higher Education, grant nr N N104 014337. A preliminary version of this paper was presented at the 4th Formulaic Language Research Network conference in Paderborn, Germany, 2010.

mainly with academic discourse (Cortes 2002; Biber et al. 2003; Biber 2004; Cortes 2004; Biber et al. 2004; Biber – Barbieri 2007; Tracy-Ventura et al. 2007). The interest in recurrent word-strings has been growing for the last decade and a half, and the method is also finding its way into historical linguistics (see Kopaczyk 2012a for an overview of ongoing and recent research).

Lexical bundles are useful for research into historical specialised discourse for three related reasons. Firstly, lexical bundles allow a "corpus-driven" rather than a "corpus-based" approach (Biber 2009: 276), which ensures objectivity. The lists of lexical bundles are products of an automatic computerised search, independent of the research question, and therefore unbiased. In other approaches to fixed expressions, for example idioms, the researcher typically comes up with an inventory on the basis of reference material and experience, and then runs searches in the database using queries based on the list of, say, idioms previously prepared (e.g. Moon 1998). This method might be useful for a study of a particular inventory. In exploring text-type fixedness, however, regardless of the expectations, the researcher will never know exactly how a specific text is structured in terms of standardized patterns, until the most characteristic, recurrent strings of lexical items are revealed. But before this happens, the researcher does not know what exact patterns are fixed and how frequently they are repeated.

It is true that after having browsed a certain number of texts the researcher gains an intuitive understanding of how the texts are structured. This kind of intuitive judgement has served researchers so far in singling out expressions characteristic for legal language. In such seminal works as Mellinkoff's *Language of the law* (1963) one finds intuitive lists of expressions typical of legal discourse in the eyes of the author, without any comment on their frequency or the actual implementation in context. Naturally, one cannot compare the robustness of research tools from fifty years ago and today. Notwithstanding this, current descriptions of legal language often repeat the traditional observations and there seems to be a gap in scholarly practice which calls for a corpus-driven, inductive methodology. This recognition has led to the employment of automatic fixed string extraction from modern legal texts (e.g. Goźdź-Roszkowski 2006). In my research, I apply corpus-driven methods to historical texts.

The corpus-driven nature of lexical bundles leads to their second asset for historical linguistics, which is the possibility of recognizing the patterns undergoing the process of standardisation. When investigating linguistic standardisation, the typical procedure would be again to select an aspect (or more) of language structure (e.g. marking plurality), check all the relevant forms in a corpus and draw conclusions. To change this approach, Biber et al. (1999) employed a simple but revolutionary method to support the choices of structures to be included in a reference grammar, or, in other words, in a model of standard struc-

tures. They searched for lexical bundles in the *Longman Spoken and Written English Corpus* (over 40mln words) and included a given string in the grammar on the basis of its frequency. In this manner, Biber et al. (1999) made a link between the frequency of occurrence on the one hand, and appropriateness on the other. In this approach, the language user becomes the source of norm, exchanging the prescriptive approach, which had always accompanied standardisation, for a descriptive approach which concentrates on the actual usage.

Thirdly, the lexical bundles method does not concentrate on phrasal constituents only, as is typically the case in other methodologies dealing with units larger than a word. The extracted strings may be smaller than a phrase (as is often the case with short bundles, 3- or 4-grams), pointing towards a certain frame which can then be filled in with various items, or they may be larger than a phrase, which then indicates the typical complementation patterns of a given phrase, or the most typical contexts in which it appears. Naturally, such findings bear significance for research into storage, processing and production issues, but this is outside the scope of the present project. What is important is that the fixing of text structure and the development of standardised patterns may not concern phrasal constituents only.

2. Standardisation on the level of text type

Even though standardisation is usually looked at in macro-scale – through the prism of prestige, formal stability and functional versatility of a given language, these aspects fall outside the scope of the present investigation.[2] Standardisation perceived in micro-scale always boils down to the selection and spread of variants on a specific linguistic level of analysis: spelling, inflections, or syntactic behaviour. The level of text is also subject to standardisation. Text types, genres and styles develop a specific repertoire of acceptable, or even expectable, constructions and phrases. For instance, studies on academic writing reach observe that it is "strongly influenced by a genre-specific standard", that "standards for academic writing are so rigid that they leave hardly any room for individual features" and that "native and non-native academics experience similar difficulties" (Locher – Strässler 2008: 13; cf. Swales 2004; Devitt 1997, 2004; and studies by Biber, Cortes and Conrad). Similarly, Bianchi and Pazzaglia (2007: 260) described a research article as a "highly standardised genre", which means that there are conventions to follow, formulae to use and specific constructions to employ, in order to construct a valid example of this genre.

---

[2] I propose to separate the extralinguistic reasons and scenarios for emerging language standards (*language standardisation*) and the language-internal processes and results of fixing specific linguistic choices (*linguistic standardisation*), see Kopaczyk 2011 and forthcoming.

Thus, for each genre or text type there will be an expected range of appropriate, acceptable constructions. The language of the text correlates with its function through prefabricated, standardised formulas and phrases, expected by the participants in a given communicative situation. Deviations from the norm may result in failing to recognise the text as the one intended, or even in a pragmatic failure, as could be the case with legal texts which may lose their force when not phrased correctly. In order to belong to a specialised discourse community, language users must be familiar with standardised patterns of creating relevant texts. As Goźdź-Roszkowski puts it, "the frequent use of [recurrent word combinations] seems to signal *competent use* within a particular register or genre, and is therefore part of 'generic knowledge' (Bhatia 1997)" (2006: 148).

Stubbs and Barth observe that recurrent word chains constitute a "predictable characteristic of different text types" and provide "evidence of units of routine language use" (2003: 62). The length of such recurrent word combinations poses very specific challenges and research opportunities. Stubbs and Barth claim that "longer chains discriminate between text types" and set the threshold for length at five elements in a string. However, they also claim that a given text type may be recognized *by* the "length of chains which recur" in it (2003: 76). In addition, Stubbs and Barth point to more formulaic types of discourse, where longer bundles are prominent, e.g. liturgical language (sermons and the Bible), political speeches and, as they hedge to admit, "some kinds of legal texts" (2003: 78). In legal discourse clarity and completeness win with stylistic variety and conscious avoidance of structural monotony, which is why legal texts are even more formulaic than other types of formal discourse. This observation has inspired the choice of long lexical bundles, eight-word chains or 8-grams, in the pilot search for standardising patterns in Middle Scots legal and administrative texts.

3. Corpora used and extraction methods

The present line of research leaves the traditional Anglo-centric mould to concentrate on the other historical national language in the island of Great Britain – Scots. The extralinguistic conditions for standardisation of Scots, as well as some aspects of language-internal standardisation, have been dealt with in detail by Devitt (1989); Leith (1997); Dossena (2003); Bugaj (2004); Millar (2005), and others. The important feature of Scots during the time frame (1380-1560) selected in the present paper is that the language was on the way towards becoming a standardised vernacular, much as its English counterpart south of the border. Uniform patterns of usage were being developed at that time, as much on other levels of linguistic analysis as on the level of text.

When it comes to legal discourse, vernacular uniformity started developing as soon as Scots took over from Latin in the legal domain. Meurman-Solin

(2000) and Rissanen (2000) suggest that legal language, in Scotland or England respectively, was a tool employed in the early stages of standardisation. Legal historians also notice the growing importance of the vernacular in Scotland. For example, the Scottish legal instrument of pleading had sources in canon law and was initially constructed in Latin. MacQueen suggests the possibility that secular forms of pleading evolved in Scotland between the thirteenth and the late fifteenth century, which "is supported by the existence of a vernacular terminology which, whatever its origins, had certainly taken on 'the protective colouring of a thoroughly native species' well before 1500" (1986: 421).

Set against this background, the current project aims to shed light on the standardisation of text in Scots legal discourse. The study is based on two complementary corpora:

1. The *Edinburgh Corpus of Older Scots* (ECOS) (2008), compiled by Keith Williamson at the University of Edinburgh. This is the newest and most comprehensive electronic corpus of medieval Scots to date, a collection of text samples from each county, designed to investigate dialectal differences in local documents; c.380,000 words, c.1380-1500.

2. The *Helsinki Corpus of Older Scots* (HCOS) (1995), compiled by Anneli Meurman-Solin at the University of Helsinki. One of the subcategories included in this multi-genre corpus includes legal and administrative localised texts from the period between 1450-1560, which amount to c.57,000 words.[3]

Altogether, the material used in the present pilot study amounts to about 450,000 words of legal and administrative Scots from 1380 to 1560. The texts were run through customised software which extracted strings of a given length – the lexical bundles – at every word boundary.[4] In Table 1, the first column lists the extracted bundle lengths, while Column 2 gives the overall count of all strings of a given length. It turns out that unique, non-repetitive bundles constitute the majority of strings (Column 3), whose ratio increases in proportion to

---

[3] Claridge (2008: 245) comments on the size of historical corpora saying that "the larger the frame, the bigger […] the corpus". Textual evidence for Scots starts in 1375 but it is only in the 15th century that the texts grow larger in numbers. For now, the ECOS coverage stops at 1500, so it was worthwhile to look for resources which would allow expanding this time frame till 1560, which is an approximate date for the imminent anglicisation of Scots. For this reason, the relevant texts from HCOS were also included in the present pilot study, while in the final version the material will be complemented with the *Wigtown Burgh Court Book* (1512-1545) (see Kopaczyk forthcoming).

[4] For a useful review of available software for lexical bundle extraction see Ari (2006). Because of the fact that the corpora used in the present project were incompatible in format and could not be easily conflated and searched, I used dedicated software written for the project by Dariusz Stróżyński.

the number of elements in a bundle. In other words, the longer the bundle, the more non-repetitive instances there are in the corpus. This result stems from the corpus-driven nature of the query. A string extracted automatically at every word boundary does not have to constitute a unit of phrasal structure or meaning. Indeed, in the majority of cases such randomly starting strings are not going to be repeated word-for-word.

Table 1. Numeric values of the searches for lexical bundles in HCOS and ECOS.

1	2	3	4	5
	all n-grams (lexical bundles)	non-repetitive (1 instance)	>1 instance	>5 instances
3-grams	559,683	329,968	229,715	8,628
4-grams	558,217	414,600	143,617	3,880
5-grams	556,751	457,740	99,011	1,830
6-grams	555,287	479,966	75,321	1,002
7-grams	553,823	492,489	61,334	552
8-grams	552,361	500,287	52,074	320

Column 4 gives all the counts for the bundles which repeat more than once. If we consider the number of 8-grams, the longest extracted strings, only about ten per cent of the total count gets repeated more than once in exactly the same wording. At this junction a question should be asked how many repeated instances would already point towards textual standardisation. Since the material generated automatically in lexical bundle queries is quite vast (Table 1), it is necessary to assume a cut-off point where for a particular corpus one can start interpreting the data (e.g. start talking about fixedness of structure and repetitiveness which links to standardisation).[5] Column 5 gives counts for bundles of a given length which repeat more than five times in an unchanged form. The search rendered 320 examples of 8-element lexical bundles employed more than five times, word for word.

It should be pointed out that the pilot searches were run on the original versions of the historical corpora, without prior spelling unification. The variety in spelling in Scots texts is immense, which constitutes the greatest obstacle in mechanical bundle extraction. The following spelling variants have been found for the most frequent 8-gram in the corpus:

[5] Scholars who employ the lexical bundle methodology set the threshold in a largely arbitrary manner. The decision depends on the research question, the length of the corpus, the length of the studied bundles, and the type of texts in the corpus (see Kopaczyk 2012a and forthcoming for an overview of cut-off points in other studies).

(1) "[be] it known to all men by these present [letters]":
IT KEND TIL AL MEN BE YIR PreseNT[6]
IT KEND TILL ALL MEn BE Yir PreseNT
IT KEND TILL ALL MEN~ BE THIR PreseNT
IT KEND TYL AL MEn BE Yir PreseNT
jT KEND TILL ALL MEN BE THIR PreseNT

A computerised search treats these strings as different instances and counts them separately. This causes a problem because some variants may in fact occur once or twice only, and the sole way to find them and include in the final count would be to search the results manually, which is not feasible especially when dealing with extremely numerous short bundles (cf. Table 1).

Any research conducted with the use of automated tools on texts dating from before the standardisation of spelling should make a provision for this complication. In order to draw lexical bundles from speech-related early modern English texts, Culpeper and Kytö (2010) unify the spelling with the help of VARD (*Variant Detector*), automatic spelling unification software developed at the University of Lancaster (Baron – Rayson 2008).[7] Kopaczyk (2012b) tests the extraction of lemmatic bundles from a lemmatised corpus. Both methods have their drawbacks. Automatic spelling unification software has been designed with English in mind, and requires a dictionary of target forms which can then be employed in form-matching algorithms. Such a dictionary for Scots would have to be produced manually (Kopaczyk forthcoming). The reliability of lemmatic bundles, on the other hand, depends on the consistency and procedure of lemmatisation in the corpus.

Having extracted 320 instances of 8-grams repeated more than five times in the raw corpus, a decision was made to unify their spelling manually. Thus, I arrived at 256 types of bundles with eight elements, of which the most frequent lexical bundle was repeated forty times, and the least frequent – five times. The most frequent of these long lexical bundles will be subject to a functional analysis below. Finally, conclusions as to their role in the creation of standardised patterns in the corpus material will be offered, together with points to consider in further research.

---

6 The ECOS uses capital letters in the transliteration of the regular script, while small font is used to represent expanded abbreviations. The HCOS does not follow this procedure. To achieve uniformity, in the data presentation below small font will be used as default because the question of expanded abbreviations is not the focus of the present investigation and the expansion does not influence the content of the lexical bundles after spelling unification.

7 For more information, see Alistair Baron's homepage http://www.comp.lancs.ac.uk/~barona/vard2/.

4. The functional categorisation of 8-grams

In their 2003 paper, Biber, Conrad and Cortes make a link between the frequency of a bundle and its importance for a specific type of discourse. "Given that lexical bundles are extremely common multi-word combinations, used widely across the texts within a register [cf. the counts in Table 1], it stands to reason that they serve fundamentally important discourse functions" (2003: 73). In this part of the paper, I aim to support this position with findings from the Scots material.

Functional categorisation, out of necessity, should rely on the most prominent function of a bundle. As Halliday put it, "[i]n general … we shall not find whole sentences or even smaller structures having just one function" (1973: 108). My categorisation is data-driven, that is it relies on the meanings and functions apparent in the most frequent bundle material in the corpus. I have based this categorisation on a broadly understood Hallidayan framework, having consulted its interpretations by Moon (1998); Cortes (2002); Biber et al. (2003, 2004); Goźdź-Roszkowski (2006); and Culpeper and Kytö (2010). Each of these projects offers a delineation of linguistic functions, ultimately stemming from Halliday's ideational, textual and interpersonal functions. Previously, I adapted this framework to the pilot study based on lemmatic bundles drawn from ECOS alone (Kopaczyk 2012b). In the present paper, I conflated the functional categories proposed by Biber, et al. (2003) and Culpeper and Kytö (2010), and arrived at three major functional groups: referential, interactional, and textual. From the 8-grams arranged according to the most prominent structural element (NP, PP, VP, or clause), I chose the top twenty bundles and analysed the discourse functions they performed. Altogether, I have categorised eighty 8-grams, according to the three major functional categories, albeit with several relevant subcategories, which were required to give credit to the specific contents of the texts.

4.1. The referential function

The referential function is parallel to Halliday's ideational function, the one related to the field. In this category I have listed the top long bundles which make reference to some aspect of the extralinguistic conditions in which the texts were constructed. Culpeper and Kytö (2010) divide the expressions from this category further, into topical and circumstantial. In the data presentation below, a semantic categorization has been applied, with such subcategories as time, location, object of legal action, and participant in legal action. Similar subdivisions were already applied in Kopaczyk (2012b) in the analysis of shorter bundles from the lemmatized version of ECOS, but some new decisions have been made since, which will be given an adequate commentary throughout the discussion below.

4.1.1. Time

Preliminary research into shorter bundles has revealed a formulaic nature of expressions introducing the day, month and year of the proceedings (Kopaczyk 2012b). It would seem that longer bundles should not emerge in this subcategory because of the diversity of the dates – suffice it to say that there were, naturally, twelve months to choose from, not to mention the number of individual years in the collection. Interestingly, however, long bundles do emerge. In these strings, the formulaic frames, identified in the search for shorter repetitive chunks, were filled with specific information which was frequent enough to be repeated in exactly the same wording several times (see examples (1)-(9) below).

1)	*of y^{e} moneth of may y^{e} zer of* 'of the month of May the year of'	17
2)	*day of y^{e} moneth of may y^{e} zer* 'day of the month of May the year'	12
3)	*god m-cccc l & vj y^{e} sutis callit* 'God 1450 and six the suits called'	12
4)	*y^{e} zer of god m-cccc l & vj* 'the year of God 1450 and six'	12
5)	*y^{e} moneth of may y^{e} zer of god* 'the month of May the year of God'	10
6)	*y^{e} zer of god m-cccc sewynti and vij* 'the year of God 1470 and seven'	10
7)	*y^{e} zer of god m-cccc sewynti and vj* 'the year of God 1470 and six'	10
8)	*of y^{e} moneth of februarij y^{e} zer of*' 'of the month of February the year of	8
9)	*of y^{e} moneth of julij y^{e} zer of* 'of the month of July the year of'	8

Several specific dates come to the fore as important in the collection and put down by the scribes in the same manner on numerous occasions. The month of May appears in examples (1), (2) and (5), while February and July are referred to in (8) and (9) respectively. These names of the months are always introduced with the formulaic sequence 'of the month of', a prepositional phrase fragment serving as a link between the day and the name of a given month. Because these months feature prominently in the texts, it may indicate an increased legal activity in burgh courts during that time. Another observation to be made about the temporal reference in long lexical bundles is the mention of God in connection with the date. As indicated in the shorter bundles (Kopaczyk forthcoming), this

is one of the possibilities, the other being 'the year of our Lord'. That expression, however, did not enter the pool of the most frequent long bundles, unlike the bundles with 'God'.

4.1.2. Location

When it comes to spatial reference, the most frequent long bundles describe the location of a plot of land undergoing a legal transaction.

10) *lyand in y^{e} burgh of peblis in y^{e}* 21
'lying in the burgh of Peebles in the'
11) *y^{e} est part on y^{e} ta syd and y^{e}* 16
'the east part on the one side and the'
12) *on y^{e} est part on y^{e} ta syd and* 16
'on the east part on the one side and'
13) *on y^{e} est part and y^{e} land of* 16
'on the east part and the land of'
14) *on y^{e} ta part and y^{e} land of* 11
'on the one part and the land of'

The idiosyncratic usage in the burgh of Peebles draws attention in example (10). No other scribe in the whole collection followed such a stable reference pattern to their burgh so frequently. The rest of the expressions is connected with delimiting land boundaries, where two competing lexemes 'part' and 'side' are employed in standardising patterns. Land issues were a major concern of the burgesses and clearly the cases connected with inheritance, sale and purchase required unchanging formulae to add validity and stability to the written confirmation of the transaction.

4.1.3. Object of legal action

The physical objects featuring in legal activity are also embedded in long fixed strings. The most prominent element in example (15) is the legal instrument, the 'letter' or 'letters', which seems to have been the most frequent expression describing the actual writ carrying official announcements, decisions and orders.

15) *til al men be yir present lettres me* 12
'to all men by this present letter me'
16) *said erd and stane in-to y^{e} handis of* 12
'said earth and stone into the hands of'

17) *y^e said erd and stane in-to ye handis* 12
'the said earth and stone into the hands'
18) *w^t erd and stane in y^e handis of* 10
'with earth and stone in the hands of'
19) *& sessyng of y^e sayd land w^t y^e* 10
'and seisin of the said land with the'
20) *of annual rent of y^e vsual mone of* 9
'of annual rent of the usual money of'
21) *of y^e said land and byggin w^t ye* 9
'of the said land and building with the'
22) *w^t y^e pertinence at twa vsual termes in* 8
'with the pertinent[s of real property] at two usual terms in'

Other objects referred to in the most formulaic strings are connected with the ritual of passing land into somebody else's hands. The *erd and stane* (examples 16-18) were literally a clod of earth and a few stones from a given piece of land, which were handed over in a small bag to the buyer or heir as a symbolic gesture of land transfer stemming from oral legal traditions (Innes 1868: xxxvii-xxxviii). Other objects of legal action referred to in a standardised manner include the land and buildings as well as tenement payments (examples 19-22).

4.1.4. Participant in legal action

In Kopaczyk (2012b) I categorized bundles making reference to the participants as expressions related to the tenor of the text. In further research, however, I have decided to keep all the references to the extra-linguistic reality, the participants and authors of the texts included, under the referential (ideational) function, and list the bundles which contain the actual activities of the participants, or the relations between them, separately under the interactional function (see 4.2. below). This separation corresponds more closely with the original Hallidayan framework, whereby the elements of text describing the participants and other components of the "organisation of experience" (Halliday 1978) are classed under the ideational function.

The long bundles indicate even longer overlapping sequences which refer to a given participant in the same manner on several occasions. Examples (23-24) pertain to the bailies of Peebles serving office at a particular point in time. There are three other contexts in which reference to bailies becomes fixed in a long lexical string. In examples (30-31), the burgh official is contextualized. An interesting trinomial made it to the top 8-grams in (34), namely *baillies counsale and communite of the burgh of*, which encapsulates the government hierarchy in a Scottish medieval burgh. A very specific formulaic reference to a sin-

gle member of the ruling elite in Peebles was made enough times in an unchanged form to be included in this discussion, see examples (36-37).

23) *of y^e balzais of peblis in yat tym* 20
'of the bailies of Peebles in that time'
24) *ane of y^e balzais of peblis in yat*t' 17
'one of the bailies of Peebles in tha
25) *twa and to y^e langar lewar of yam* 13
'two and to the [one living longer] of them'
26) *his spous and gaif to yam twa and* 12
'his spouse and gave to them two and'
27) *spous and gaif to yam twa and to* 12
'spouse and gave to them two and to'
28) *to yam twa and to y^e langar lewar* 12
'to them two and to the [one living longer]
29) *yam twa and to y^e langar lewar of* 12
'them two and to the [one living longer] of'
30) *balze in y^t tym and yan in-continent ye* 12
'bailie in that time and then immediately the'
31) *hand balze in y^t tym and yan in-continent* 12
'hand bailie in that time and then immediately'
32) *notar and common vritar of y^e said burgh* 12
'notary and public scribe of the said burgh'
33) *public notar and common vritar of y^e said* 12
'public notary and public scribe of the said'
34) *balzais counsale and communite of y^e burgh of* 10
'bailies council and community of the burgh of'
35) *of crawfurde public notar and common vritar of* 9
'of Crawford public notary and public scribe of'
36) *in-to y^e handis of partick dikeson ane of y^e* 8
'into the hands of Patrick Dickson one of the'
37) *of partick dikeson ane of y^e balzeis od peblis* 8
'of Patrick Dickson one of the bailies of Peebles'

Another long sequence of overlapping standardised strings is found in the context of inheritance after the death of a spouse (25-29). In its full version, the fixed pattern of reference connects the act of giving and the person who receives the rights to the inherited possessions: *his spous and gaif to yam twa and to ye langar lewar of yam*. In this structurally complex fixed string, *his spous* is a continuation of the preceding clause, which must have ended that clause on enough occasions to make it to the most repetitive pool of long bundles. The

lexical bundles often point to cross-clausal and cross-phrasal fixedness, which is impossible to discover with a different methodology.

In the bundles we find explicit reference to the person putting down documents and court proceedings, also in the form of a binomial: *(public) notar and common writar* (32-33). Interestingly, none of the hereditary relationships revealed in the shorter bundles search, e.g. *heir and assignee* or *executor and assignee* (Kopaczyk 2012b), have emerged as part of the longer fixed strings.

4.2. The interactional function

This functional category includes bundles which represent the action taking place between the participants of a communicative situation. Culpeper and Kytö (2010: 110) call this category 'interpersonal' and apply it to speech acts and modal meanings. The interactional category in Biber et al. (2003) is separated from stance expressions, which I nevertheless blend together here, because hedges and qualifiers (see section 4.2.5. below) arise in interaction and add to a particular contextual understanding of an activity. This category also includes the expanded versions of shorter bundles classified in Kopaczyk (2012b) under: "Legal action: Focus on the court" and "Legal action: Focus on the people".

4.2.1. Directives

Within the group of interactional bundles, the most prominent are directives, where the court or another authority passes an order, makes a law known and binding or calls a case. The most frequent 8-grams in the material from ECOS and HCOS are concerned with the force of the legal instrument and add up to an extended directive sequence "(be) it known to all men by the present (letter)" (38-40). The numeric discrepancies between the overlapping bundles indicate that the core of a given bundle may have remained unchanged, while the preceding and following co-text may have fluctuated. For example, it is possible that the prepositional phrase fragment at the end of (39) would be complemented only with 'letters' on several occasions, without the premodifier 'present', or even with a synonym to the noun 'letters'.

38) *it kend til all men be yir present* 55
'it known to all men by this present'
39) *be it kend til all men be yir* 38
'be it known to all men by this'
40) *kend til all men be yir present lettres* 28
'known to all men by this present letter'

41) *zerly to be rasit and takyn of ye* 9
'yearly to be raised and taken from the'
42) *it is statut and ordanit be the quenis* 8
'it is stated and ordained by the Queen's'
43) *is statut and ordanit be the quenis grace* 8
'is stated and ordained by the Queen's grace'
44) *item it is statut and ordanit that the* 8
'likewise it is stated and ordained that the'
45) *ordanit be my lord gouernor with auise of* 8
'ordained by my Lord Governor with advice of'
46) *it is deuisit statut and ordanit be my* 7
'it is devised stated and ordained by my'
47) *it is statut and ordained that the act* 7
'it is stated and ordained that the act'
48) *& his party sal be vnscathit of him* 7
'and his share shall be unharmed by him'
49) *his party sal be vnscathit of him* 7
'his share shall be unharmed by him and'
50) *party sal be vnscathit of him* 7
'share shall be unharmed by him and his'

There are several binomial expressions embedded in the long directive bundles, e.g. *rasit and takyn* (41), *statut and ordanit* (42-44, 47), and also one trinomial *deuisit statut and ordanit* (46). The formulaic binomial "stated and ordained" has two standard complementation patterns, which get revealed in the search for longer bundles: a PP in *by* and a relative clause.

4.2.2. Representatives

Representatives "commit the speaker to the truth of the expressed proposition" (Levinson 1983: 240). This group of long repetitive strings includes standardised ways in which the court confirms or affirms a given state of affairs or a result of the legal activity.

51) *ye-quhilk day y^{e} sutis callit y^{e} curt effermit* 13
'the which day the suits [were] called [which] the court affirmed'
52) *y^{e} sutis callit y^{e} curt effermit ilk absent* 12
'the suits [were] called [which] the court affirmed each absent [one]'
53) *sutis callit y^{e} curt effermit ilk absent in* 12
'suits [were] called [which] the court affirmed each absent [one] in'

It turns out that the most standardised confirmations (51-53) have to do with absences which were a notorious problem in medieval and early modern Scottish burgh courts. The burgesses were obliged by law to take part in the burgh court and council. The court kept record of the absences, so that a fine or some other type of punishment was ordained with the third absence.

4.2.3. Declaratives

Declaratives can be rephrased by means of a *hereby*-structure and "effect in immediate changes in the institutional state of affairs and […] tend to rely on elaborate extra-linguistic institutions" (Levinson 1983: 240).

54) *m-cccc l & vj y^{e} sutis callit y^{e}* 12
'1450 and six the suits [were] called the'
55) *past w^{t} y^{e} said lettres & yir witnes* 10
'passed with the said letters and this witness'
56) *j past w^{t} y^{e} said lettres & yir* 10
'I passed with the said letters and this'
57) *l & vj y^{e} sutis callit y^{e} curt* 6
'50 and six the suits [were] called the court'

The standardised patterns in this group refer to passing a new law or obligation by means of a writ (55-56), as well as to the calling of suits at court (54, 57). In shorter bundles, this category should be more prominent because the variable contextualization (e.g. the year) would not interfere with the structure of an extracted string.

4.2.4. Commissives

In general, commissives include speech acts which commit the speaker to a particular course of action, and often contain such verbs as *promise*, *vow* or *swear* (Levinson 1983). In examples (58-60), commitment is implied but not verbalized overtly. By giving permission in (58), the sender of the message commits himself to taking some further steps (the details are missing from the standardising part). In (59-60), the advice of the three Estates of the Scottish parliament may also be perceived as a commitment to endorse a particular course of action.

58) *is w^{t} consent of party continuit to y^{e}* 12
'is with consent of [the] party continued to the'

59) *with auise of the thre estatis of parliament* 15
'with advice of the three Estates of Parliament'
60) *auise of the thre estatis of parliament* 13
'advice of the three Estates of Parliament that'

It is worth mentioning that burgesses were represented in the Parliament as one of the aforementioned estates, which clearly emphasises their position and importance in the administration and government of the kingdom.

4.2.5. Hedges and qualifiers

The last examples in the interactive category qualify a given proposition, either by attributing it to God (61) or by explaining the reason (62).

61) *be y^{e} grace of god king of scottis* 9
'by the grace of God king of Scots'
62) *of my i curt for falt of entres* 9
'of my first court for fault of entry'

The phrasal core of the long repetitive chunk in (62), i.e. "for fault of entry", in fact surfaced in the overall search for most formulaic lemmatic bundles in ECOS, as well (Kopaczyk 2012b).

4.3. The textual function

This is the group of expressions which create the text reality, make cohesive connections within it and relate the text to other texts (see also Biber et al. 2003). Following Culpeper and Kytö (2010: 110), it was decided that this separate group of bundles will not be limited to various aspects of cohesion but it will group the elements of narration which emerge in a fixed format. As in the case of other long bundles, a given string rarely carries a single communicative function, which often makes it difficult to categorize the strings. Textual 8-grams contain elements of description, references to extra-linguistic reality, or parts of speech acts, which is why some of their overlapping continuations were already discussed above (cf. examples 16-18 and 25-29). As outlined in the methodology section, however, I based the division into different functions on the most prominent functional element in a long string, such as grammatically and semantically complete cores. Thus, *yam twa* and *ye langar lewar* constitute core elements in the string *to yam twa and to ye langar lewar* (28), which places this bundle in the referential function, while in the string *gaif to yam twa and to ye langar* (65), the core element is the verb 'gave', which places it in the narrative subsection in the textual function.

4.3.1. Narrative

The bundles in this subgroup contain an element of narration, a sequence of events conveyed by action verbs. In fact, three court activities emerge in the form of long lexical strings which overlap to a significant extent.

63) *y^e said balze deliuerit and laid y^e said* 14
'the said bailie delivered and laid the said'
64) *and gaif to yam twa and to y^e* 13
'and gave to them two and to the'
65) *gaif to yam twa and to y^e langar* 13
'gave to them two and to the'
66) *deliuerit and laid y^e said penny in-to y^e* 9
'delivered and laid the said penny into the'
67) *and laid y^e said penny in-to y^e handis* 9
'and laid the said penny into the hands'
68) *laid y^e said penny in-to y^e handis of* 9
'laid the said penny into the hands of'
69) *deliuerit and laid y^e said erd and stane* 9
'delivered and laid the said earth and stone'
70) *balze deliuerit and laid y^e said erd and* 8
'bailie delivered and laid the said earth and'
71) *laid y^e said erd and stane in-to y^e* 8
'laid the said earth and stone into the'
72) *said balze deliuerit and laid y^e said erd* 7
'said bailie delivered and laid the said earth'
73) *in-continent y^e said balze deliuerit and laid y^e* 6
'immediately the said bailie delivered and laid the'

The first overlapping sequence adds up to *in-continent y^e said balze deliuerit and laid y^e said penny in-to y^e handis of* (examples 63, 66-68, 73), which refers to the ritual of transferring land with an in- and out-penny (Innes 1868: xxxvii-xxxviii). The lexical bundle method reveals a predisposition for this event to be recorded in the same wording on several occasions. Interestingly, the initial part of the bundle may have been shared with another long lexical string, which is very similar in meaning: *in-continent y^e said balze deliuerit and laid y^e said erd and stane* (examples 69-72). As explained above, earth and stones from a plot under transaction were used in a symbolic gesture of transfer performed in the legal process. Finally, the third narrative emerging from the bundles concerns the action of giving something, most probably money or goods – perhaps too many options were possible here to qualify into the pool of long bundles – to the remaining spouse (examples 64-65).

4.3.2. Definitional

The final subcategory includes strings which perform a definitional function. On the one hand, the text aims at clarifying preceding information, using such expressions as 'after + N' (74), ‘that is to say’ (75) and ‘as follows thereafter’ (76-77).

74) *in all punctis efter y^{e} forme and tenour* 9
‘in all points after the form and tenor’
75) *men y^{t} is to say jhon off kynharde* 8
‘men that is to say John of Kinhard’
76) *& effect as efter followis that is to* 6
‘and effect as follows thereafter that is to’
77) *effect as efter followis that is to say* 6
‘effect as follows thereafter that is to say’

On the other hand, the definitional function may also be carried through reference to other texts, that is through intertextuality (78-80).

78) *& effect as efferis and as it was* 10
and effect as is suitable and as it was’
79) *effect as efferis and as it was ye* 10
‘effect as is suitable and as it was the’
80) *as efferis and as it was ye said* 10
‘as is suitable and as it was the said’

According to the DSL, the expression *as efferis* is a “formal phrase” with the meaning “as is suitable”, from OF *aferir* ‘to belong, to pertain’. What “is suitable” in a legal context relies heavily on traditions and earlier laws, so this lexical bundle sets the current state of matters against the earlier legal background, making this function intertextual.

5. Concluding remarks

Long lexical bundles offer a unique insight into textual fixedness and standardisation. The findings prove that even 8-grams are repetitive – the extraction rendered 320 different types of 8-grams, which repeated in fixed form over five times in a corpus of c.450,000 words, and ranged from five to fifty five instances of a given string. Word-to-word repetition of eight elements in the same sequence – in relatively large numbers in a corpus of such a size – indicates the existence of formulaic, usual patterns and standardising ways of phrasing some important meanings.

The division of the extracted material into functional categories draws attention to text type-specific meanings and functions. Tannen (1987) saw language as a 'cultural encoder', concerned with "ideas that are familiar to the language community, with how things are commonly said in that community...". The burgh community of medieval and early modern Scotland produced legal discourse which answered to its communicative needs within this particular sphere of life. What comes to the fore in the form of long lexical bundles is the referential function, especially in reference to the participants of legal proceedings. Other important standardising areas were time and space dimensions, as well as typical reference to the objects involved in legal activity. It seems that the scribes were likely to phrase these references in a fixed manner, or at least they were sensitive to emerging conventions.

In terms of frequency of repetition, the most frequent individual bundles can be found among directives, which are tokens of the interactional function and the performative nature of language. In a legal context, the extralinguistic reality is, indeed, shaped through direct and indirect speech acts. Speech act-related bundles occur frequently in a standardised form in the extracted material, which may happen because the felicity conditions require the wording to be formulaic and stable. This fixing of form can point towards the recognition that only in the same format can the same speech act be valid throughout the records.

Long lexical bundles, unlike their shorter versions, are rarely, if ever, built out of complete grammatical constituents or functional elements. This is why the bundles containing features of cohesion and narration overlapped with other bundles, for example referential. Generally speaking, syntagmatic overlaps[8] are a pervasive feature of long bundles, pointing towards even longer patterns in the texts. In individual cases, such repetitive strings may consist of even more than ten lexical elements.

On a concluding note, it is worth stressing that the long bundles extracted from legal and administrative historical Scots records were all text-type specific. This assertion supports the suggestion that long strings repeated in the same form are required by specialised discourse in the area of law, which was put forward tentatively by Stubbs and Barth (2003) and discussed in section 2. In further research it will be worthwhile to trace chronological and geographical tendencies for the development of textual patterns. A new understanding of the degree and kind of fixing and patterning in legal discourse may also come from the comparison of short and long bundles from the same collection of texts (see Kopaczyk forthcoming).

---

[8] Syntagmatic overlaps are linear, which means that some part of a given bundle becomes part of another bundle, see, e.g., examples (78-80). I distinguish such overlaps from *paradigmatic overlaps*, which can be seen when a shorter repetitive string is included within a longer, also repetitive frame, and is exchangeable for some other unit (Kopaczyk forthcoming).

REFERENCES

PRIMARY SOURCES

Meurman-Solin, Anneli (ed.)
1995 *The Helsinki Corpus of Older Scots* [HCOS]. Helsinki: University of Helsinki.
Williamson, Keith
2008 *The Edinburgh Corpus of Older Scots* [ECOS]. Edinburgh: University of Edinburgh.

SECONDARY SOURCES

Ari, Omer
2006 "Review of three software programs designed to identify lexical bundles", *Language Learning and Technology* 10/1: 30-37.
Baron, Alistair – Paul Rayson
2008 "VARD 2: A tool for dealing with spelling variation in historical corpora", *Proceedings of the Postgraduate Conference in Corpus Linguistics*, Aston University, Birmingham, 22 May 2008. <http://acorn.aston.ac.uk/conf_proceedings.html>
Bianchi, Francesca – Roberto Pazzaglia
2007 "Student writing of research articles in a foreign language: metacognition and corpora", in: Roberta Facchinetti (ed.), 259-287.
Biber, Douglas
1997 "Lexical bundles in spoken and written discourse: What the grammar books don't tell you", in: S.B. Gerome (ed.), 4-8.
Biber, Douglas
2004 "Lexical bundles in academic speech and writing", in: Barbara Lewandowska-Tomaszczyk (ed.), 165-178.
Biber, Douglas
2009 "Corpus-based and corpus-driven analyses of language variation and use", in: Bernd Heine – Heiko Narrog (eds.), 159-191.
Biber, Douglas – Francesca Barbieri
2007 "Lexical bundles in university spoken and written registers", *English for Specific Purposes* 26: 263-286.
Biber, Douglas – Susan Conrad – Viviana Cortes
2003 "Lexical bundles in speech and writing: An initial taxonomy", in: Andrew Wilson et al. (eds.) 71-92.
Biber, Douglas – Susan Conrad – Viviana Cortes
2004 "If you look at... Lexical bundles in university teaching and textbooks", *Applied Linguistics* 25/3: 371-405.
Biber, Douglas – Stig Johansson – Geoffrey Leech – Susan Conrad – Edward Finegan
1999 *Longman grammar of spoken and written English*. London: Longman.
Bishop, Wendy – Hans Ostrum (eds.)
1997 *Genre and writing*. Portsmouth: Boynton / Cook.
Bugaj [Kopaczyk], Joanna
2004 "Middle Scots as an emerging standard and why it did not make it", *Scottish Language* 23: 19-34.

Claridge, Claudia
2008 "Historical corpora", in: Anke Lüdeling – Merja Kytö (eds.), 242-258.
Conrad, Susan – Douglas Biber
2004 "The frequency and use of lexical bundles in conversation and academic prose", *Lexicographica: International Annual for Lexicography* 20: 56-71.
Cortes, Viviana
2002 "Lexical bundles in freshman compositions", in: Randi Reppen et al. (eds.), 131-146.
Cortes, Viviana
2004 "Lexical bundles in published and student writing in history and biology", *English for Specific Purposes* 23/4: 397-423.
Culpeper, Jonathan – Merja Kytö
2010 *Early Modern English dialogues: Spoken interaction as writing*. Cambridge: Cambridge University Press.
Deumert, Ana – Wim Vandenbussche (eds.)
2003 *Germanic standardizations. Past to present*. Amsterdam – Philadelphia: John Benjamins.
Devitt, Amy
1989 *Standardising Written English. Diffusion in the Case of Scotland 1520-1659*. Cambridge: Cambridge University Press.
Devitt, Amy
1997 "Genre as a language standard", in: Wendy Bishop – Hans Ostrum (eds.), 45-55.
Devitt, Amy
2004 *Writing genres*. Carbondale: Southern Illinois University Press.
Dossena, Marina
2003 "Scots", in: Ana Deumert – Wim Vandenbussche (eds.), 383-404.
Dictionary of the Scots Language [DSL] <www.dsl.ac.uk>
Facchinetti, Roberta (ed.)
2007 *Corpus linguistics 25 years on*. Amsterdam: Rodopi.
Gerome, S.B. (ed.)
1997 *An update on grammar: How it is learnt – How it is taught* (1996 Colloquium Proceedings). Paris: TESOL France.
Gotti, Maurizio – Davide Giannoni (eds.)
2006 *New Trends in Specialized Discourse Analysis.* Frankfurt a/Main: Peter Lang.
Goźdź-Roszkowski, Stanisław
2006 "Frequent phraseology in contractual instruments: A corpus-based study", in: Maurizio Gotti – Davide Giannoni (eds.), 147-162.
Heine, Bernd – Heiko Narrog (eds.)
2009 *The Oxford handbook of linguistic analysis*. Oxford: Oxford University Press.
Innes, Cosmo
1868 "Preface", in: Cosmo Innes (ed.), xx-l.
Innes, Cosmo (ed.)
1868 *Ancient laws and customs of the burghs of Scotland*. Vol. 1. *A.D.1124-1424*. Edinburgh: Scottish Burgh Records Society.
Halliday, M.A.K.
1973 *Explorations in the functions of language*. London: Edward Arnold.
1978 *Language as a social semiotic*. London: Edward Arnold.

Kopaczyk, Joanna
2011 "Standaryzacja tekstów w perspektywie historycznej. Analiza zbitek leksykalnych" [Text-type standardisation in a historical perspective. Analysing lexical bundles], in: Piotr Stalmaszczyk (ed.), 155-174.
2012a "Applications of the lexical bundles method in historical corpus research", in: Piotr Pęzik (ed.).
2012b "Repetitive and therefore fixed? Lemmatic bundles and text-type standardisation in fifteenth-century administrative Scots", in: Hans Sauer – Gaby Waxenberger (eds.) Forthcoming *Standardising legal discourse. The language of Scottish burghs 1380-1560*. (Language and the law 1. Series Editor: Roger W. Shuy). Oxford University Press.
Leith, David
1997 *A social history of English* (2nd edition). London: Routledge.
Lewandowska-Tomaszczyk, Barbara (ed.)
2004 *Practical applications in language corpora (PALC 2003).* Hamburg: Peter Lang.
Levinson, Stephen C.
1983 *Pragmatics*. Cambridge: Cambridge University Press.
Locher, Miriam A. – Jürg Strässler
2008 "Introduction: Standards and norms", in: Miriam A. Locher – Jürg Strässler (eds.), 1-20.
Locher, Miriam A. – Jürg Strässler (eds.)
2008 *Standards and norms in the English language*. Berlin – New York: Mouton de Gruyter.
Lüdeling, Anke – Merja Kytö (eds.)
2008 *Corpus linguistics. An international handbook* (vol.1). Berlin – New York: Walter de Gruyter.
MacQueen, Hector L.
1986 "Pleadable brieves, pleading and the development of Scots law", *Law and History Review* 4/2: 403-422.
Mellinkoff, David
1963 *The language of the law*. Boston, Toronto: Little, Brown and Company.
Meurman-Solin, Anneli
2000 "Change from above or from below? Mapping the loci of linguistic change in the history of Scottish English", in: Laura Wright (ed.), 155-170.
Millar, Robert McColl
2005 *Language, nation and power*. Basingstoke: Palgrave Macmillan.
Moon, Rosamund
1998 *Fixed expressions and idioms in English*. Oxford: Clarendon Press.
Parodi, G.iovanni (ed.)
2007 *Working with Spanish corpora.* London: Continuum.
Pęzik, Piotr (ed.)
2012 *PALC Proceedings*. Frankfurt a/Main: Peter Lang.
Reppen, Randi – Susan M. Fitzmaurice – Douglas Biber (eds.)
2002 *Using corpora to explore linguistic variation*. Amsterdam: John Benjamins.
Rissanen, Matti
2000 "Standardisation and the language of early statutes", in: Laura Wright (ed.), 117-130.

Sauer, Hans – Gaby Waxenberger (eds.)
2012 *Papers from the 15 ICEHL*. Frankfurt a/Main: Peter Lang.
Stalmaszczyk, Piotr (ed.)
2011 *Metodologie językoznawstwa. Od ontologii do pragmatyki* [Linguistic methodology. From ontology to pragmatics]. Łódź: Wydawnictwo Uniwersytetu Łódzkiego.
Stubbs, Michael – Isabel Barth
2003 "Using recurrent phrases as text-type discriminators: A quantitative method and some findings", *Functions of Language* 10/1: 61-104.
Swales, John M.
2004 *Research genres. Exploration and application*. Cambridge: Cambridge University Press.
Tannen, Deborah
1987 "Repetition in conversation as spontaneous formulaicity", *Text* 7: 215-243.
Taavitsainen, Irma – Gunnel Melchers – Päivi Pahta (eds.)
1999 *Writing in Nonstandard English*. Amsterdam: John Benjamins,
Tracy-Ventura, Nicole – Douglas Biber – Viviana Cortes
2007 "Lexical bundles in Spanish speech and writing", in: Giovanni Parodi (ed.), 217-231.
Wilson, Andrew – Paul Rayson – Tony McEnery (eds.)
2003 *Corpus linguistics by the Lune. A Festschrift for Geoffrey Leech*. Frankfurt a/ Main – New York: Peter Lang.
Wright, Laura (ed.)
2000 *The development of Standard English, 1300-1800*. Cambridge: Cambridge University Press.