Formulaic discourse across Early Modern English medical genres

Investigating shared lexical bundles*

Joanna Kopaczyk
Adam Mickiewicz University

This paper offers a corpus-driven investigation into the formulaic nature of Early Modern English medical genres. The aim of this study is to answer three related questions: (1) to what extent various text categories in medical discourse share the same lexico-syntactic choices?; (2) what stable and fixed lexico-syntactic patterns repeat across various texts related to medicine?; and (3) is there a diachronic dimension to the employment of these repetitive strings? The study is based on the recently published electronic corpus of *Early Modern English Medical Texts* (EMEMT, 1500–1700, Taavitsainen et al. 2010) and uses the lexical bundle method (Biber et al. 1999) to extract 3-grams from the normalized version of the corpus. The diachronic distribution of 3-grams across medical texts shows an increase in the number of text categories which share lexical bundles. When it comes to specific 3-grams, the paper presents a diachronic overview of the most prominent semantic areas where overlaps of fixed strings occur among text categories, e.g. quantification, body parts, time and sequence, or ingredients. The study has also found important overlaps in purely functional contexts, e.g. in clarification, modality or efficacy expressions, and in structural frames, e.g. copula constructions and prepositional phrase fragments. With the help of an objective, frequency-driven corpus tool, the common lexico-syntactic core of early modern medical discourse could be established. At the same time, clusters of text categories sharing the same preferences could emerge.

1. Introduction: Formulaic discourse in historical texts

Different genres require different degrees of linguistic creativity, depending on their aim and function. For instance, those which aim to entertain the readers or

* I would like to acknowledge the support of the Swiss National Foundation, which helped me to attend the ICEHL17 conference. This paper has benefitted from the comments offered by the audience and from the insightful suggestions made by an anonymous reviewer.

sell them a product may contain a large degree of linguistic innovation, surprising collocations, or one-off metaphoric phrases. At the same time, in those genres, where the stability of form creates reliability and authority of the text, e.g. in the legal context or in religious rituals, the amount of linguistic innovation will be suppressed and more structures will be repeated in the same form, without looking for alternative ways of expressing a given meaning (for a thorough discussion of legal language see Danet 1980; Tiersma 1999, 2006; on religious discourse see Kohnen 2010). This genre-related requirement of textual fixedness goes hand in hand with the idiom principle, formulated by Sinclair (1991: 110): "a user has available to him a large number of semi-preconstructed phrases that constitute single choices, even though they might appear to be analysable into segments". The "single choices" – the fixed and repetitive configurations of lexemes – are part of our linguistic capacity in general but they may be activated to a larger degree when we construct texts of a particular kind, be it in speech or in writing. The more structures from a given text repeat across similar texts, the more formulaic the genre is to which these texts belong.

It has been pointed out by Taavitsainen (2010: 32) that genres "provide important clues to meaning-making practices, and they change according to the needs, goals and likings of their users". Genres are therefore dynamic entities whose features evolve in time, and it is crucial to consider what may be subject to change and what is stable when one wants to understand the construction of meaning in a given register, or discipline, or textual tradition. Repetitive and unchanging fragments of the textual fabric in a given genre convey important meanings for the audience and create a stable framework for the communicative functions of the text. The form, content, and placement of these stable fragments may be subject to change within a given genre, in line with Taavitsainen's suggestion, which makes diachronic investigations of this issue even more pertinent.

Historical linguists have recently started looking into the formulaic nature of genres from the past, making use of electronic corpora and automatic search methods (see Section 2.2). Following the publication of the electronic corpus of *Early Modern English Medical Texts* (EMEMT, 1500–1700, Taavitsainen et al. 2010), it is now possible to engage in such research on the basis of early modern medical texts and explore repetitive patterns in this type of discourse. In the manual to the dedicated software, EMEMT compilers showcase the features of the corpus which should be useful in investigating formulaic structures and encourage scholars to take up this line of study (Tyrkkö, Hickey, & Marttila 2010: 258). Thus, the present paper aims to investigate the degree and character of formulaicity in early modern medical texts, on the basis of EMEMT,

a representative and comprehensive corpus of such texts written in English in the sixteenth and seventeenth centuries. More specifically, a question to be answered is what structures repeat across the different types of texts classified as medical and included in the corpus, and whether these repetitive structures constitute some kind of linguistic core of English medical discourse of the day. In order to investigate these issues in a reliable manner, I have selected a corpus-driven method of data extraction, the lexical bundles.

2. Lexical bundles in historical research

2.1 Introduction to lexical bundles

A corpus methodology known as *lexical bundles*[1] was developed in the mid-90s, with the first studies into repetitive strings in spoken English (e.g. Altenberg 1998). The extraction of lexical bundles consists in running the electronic text through software which automatically identifies repetitive strings of words of a given length, *n*-grams (e.g. 3-grams, see Section 3.2), regardless of their semantic or structural completeness. The monumental and innovative *Longman Grammar of Spoken and Written English* (Biber et al. 1999) used this method to identify language user preferences in grammar, as well as areas of fixedness in different discourse situations, on the basis of robust present-day corpora of spoken and written English. The definition of lexical bundles found in Biber et al. (1999) stresses the recurrent quality and high frequency of bundles, repeated in a corpus in exactly the same form. As this method is corpus-driven rather than corpus-based (Tognini-Bonelli 2001: 84–87), the search results are not geared towards any particular research question. It is the researcher who chooses to interpret the results in a given manner, while the same pool of lexical bundles may serve different investigations. One of the study areas in which lexical bundles are helpful is formulaicity research because the bundles capture the elements of discourse which are stable and repetitive and, therefore, must fulfill an important function in the text (cf. Biber 2009: 282–286).

1. The same kind of structures has also been labelled as word clusters (Scott 1997), *n*-grams, repetitive word strings, recurrent word chains (Stubbs & Barth 2003), contiguous formulaic strings (Conklin & Schmitt 2008), or lexical clusters (EMEMT).

As an illustration of this method and the potential functional interpretation of its results, consider these fragments of classroom conversations and textbooks (Biber, Conrad & Cortes 2004):

Table 1. A functional categorization of 4-grams extracted from the *T2K-SWAL Corpus* (based on Biber, Conrad & Cortes 2004)

	Discourse organisers	Attitudinal stance
spoken	*what do you think*	*if you want to*
	if you have a	*I don't want to*
	if we look at	*do you want to*
	want to do is	*you don't have to*
	going to talk about	*do you want me*
	has to do with	*going to have to*
written	*in this chapter we*	*it is important to*
	on the other hand	*it is necessary to*
	as well as the	*can be used to*

As the material in Table 1 clearly shows, lexical bundles are not complete phrases but rather phrase fragments, which opens an interesting perspective on linguistic fixedness outside the traditionally defined complete structural units. The bundles above have been grouped according to functions in a Hallidayan tradition: the discourse or the textual function in the first column, and the stance or modality function in the second column of Table 1 (see Halliday 1978 for the original conceptualization of the semanto-functional components of language: ideational, interpersonal, and textual). These bundles may also be interpreted as formulaic elements of academic discourse, repeated in expected contexts in an unchanged form. Comparing the frequencies of selected strings may indicate which of these bundles are more formulaic than others. The same methodology may be employed in the study of historical medical texts to explore their formulaic ingredients.

2.2 Lexical bundles in historical corpora

The lexical bundle method has been adapted to other areas of linguistic research, also in a historical context.[2] In their survey of variation and change

2. For a review of potential applications of the lexical bundles in historical linguistics, and the methodological problems involved in using this method, see Kopaczyk (2012a). The present section is an extended summary of the main points put forward in that publication.

in nineteenth-century English, which employs lexical bundles alongside other corpus methods, Kytö and Smitterberg (2006: 200) notice that "[t]he occurrence of lexical bundles in Present-day English has received a great deal of attention in recent years, but not much is known about their distribution in historical texts". So far, two major projects have adapted lexical bundles to working with historical corpora from before 1700: Culpeper and Kytö (2010: 103; see also their pilot paper, Culpeper & Kytö 2002) used lexical bundles to "investigate the role played by recurrent word-combinations in speech-related language of the Early Modern English period" while Kopaczyk (2013) extracted lexical bundles from the fifteenth- and sixteenth-century administrative and legal texts written in Scots to trace patterns of textual standardization. One of the reasons why this method is difficult to apply in historical linguistic research is spelling variation, which impedes reliable automatic extraction of identical strings and distorts the results of automatic queries (see Section 2.3 for illustration). Consequently, drawing bundles from medieval or early modern texts requires methodological caution and artificial spelling normalization.

Solutions to the problem of spelling variation are based on automatic spell-checker algorithms and calculations of distance between different arrangements of characters. The tool called VARD (Variant Detector), developed at Lancaster University (for the homepage see Baron 2010; also Rayson et al. 2007; Baron & Rayson 2008; Lehto et al. 2010) is a pioneer in this area and has been designed to work with Early Modern English texts. Its more recent versions can be trained to normalize spelling in other texts after uploading a dictionary of target forms and specifying the replacement rules. Producing this kind of a dictionary for language-states with extensive spelling variation, such as Middle English, might be problematic in itself, so VARD, a useful tool as it is, cannot be readily applied to all types of corpora.

Another problem in using automatic extraction methods in historical corpora is caused by the lack of uniformity in digitizing conventions and editorial intervention. In every corpus or electronic text edition, the editors take their own decisions how to represent the reality of handwritten or early printed texts (Robinson 2009).[3] For instance, some corpora may be based on edited texts, which have silently expanded abbreviations, while others may use italics to indicate abbreviated sequences or choose not to expand abbreviations and represent

3. For illuminating discussions of the rationale behind transcription conventions and representing the manuscript spelling reality, see the contributions in Blake and Robinson (eds 1993, *The Canterbury Tales Project*).

them with a symbol. It is true that modern corpora include information on the adopted editorial practices (Kytö 2012: 1513–1514), but it does not mean that all electronic editions are prepared according to the same principles. This may become an issue when one wants to draw lexical bundles from combined corpora or from digital repositories based on varied material in terms of medium (manuscripts, *incunabula*, later prints) or perform comparative analyses across different corpora.

The extraction software itself may pose problems too. There are open-source programs for extracting lexical bundles but not all of them can handle historical data or corpus mark-up. Ari's (2006) test showed that different software packages may render different results when used on the same corpus. Since then, no similar investigation has been carried out but one may suspect that the discrepancies persist.

Finally, historical corpora are generally much smaller than the robust present-day language repositories. Textual sources have been limited in an ad hoc manner in the course of time and not all surviving texts are searchable automatically. In addition, stratified historical corpora comprising, for instance, a specific type of discourse or a selected historical period (Kytö 2012: 1510), may be relatively small, e.g. under one million words. Smaller size does not impede the usefulness of a corpus, as Kytö (2012: 1516–1517) recently argued. However, in the case of lexical bundles, a small size of the corpus may have influence on the cut-off point, where a scholar decides, in a rather arbitrary manner (cf. Biber & Barbieri 2007: 267), how many instances of a given fixed and recurrent string will mean that the string qualifies as a lexical bundle (for a comparison of cut-off points in ten studies based on lexical bundles see Kopaczyk 2013: 152–153).

2.3 Lexical bundles: Solutions for EMEMT

The compilers of the EMEMT corpus (Taavitsainen et al. 2010) have given careful consideration to the points raised above. What is especially valuable is that the question of spelling normalization was addressed and that the VARD 2.3. software was used to produce a normalized version of the corpus files (on adapting VARD for EMEMT see Lehto et al. 2010). Thus, the researcher may choose a regular version of the corpus or a normalized version, which would be crucial for the extraction of lexical bundles. As explained in the previous section, drawing lexical bundles from a corpus with no standardized spelling distorts the findings. Consider the Examples (1a–b), selected from surgical treatises (Category 5):

(1) a. EMEMT, regular files, Category 5 (3-grams, 5 tokens and above)
according to the 69 tokens
accordyng to the 9 tokens
accordynge to the 10 tokens

b. EMEMT, normalized files, Category 5 (3-grams, 5 tokens and above)
according to the 90 tokens

The same search settings render different results for the regular and for the normalized files: in the first case, the same lexical bundle was counted separately three times due to its varied spelling; in the second case, the bundle had a higher token score, which incorporated all the variants. In fact, a careful reader will notice that since the search in the normalized version returned 90 tokens, two tokens are missing. A manual search through the regular version found these two missing instances: the spelling *accordinge to the* was used twice, so it did not qualify above the cut-off point and was missed out in the first search for lexical bundles. In EMEMT, the processing of regular and normalized files is performed by EMEMT Presenter, a special customized version of the Corpus Presenter software package, developed by Raymond Hickey.[4] The close cooperation between the software author and the corpus compilers (cf. Tyrkkö, Hickey & Marttila 2010) has boosted the reliability of dedicated automatic tools for bundle extraction.

The problem of various digitizing conventions and editorial inconsistencies is not present here either because the corpus transcriptions are consistent among the parts and thoroughly explained in the contributions on individual text categories in Taavitsainen and Pahta (eds 2010). The size of the corpus is substantial for a repository of historical texts (c. 2mln words) and the cut-off points are adjustable, which allows a researcher to create different queries and compare their results (on the methodological decisions taken in the present study see Section 3.2).

3. Formulaic language in medical texts

3.1 Previous scholarship

The notion of formulaicity permeates the discussion of medical discourse, but it is usually highlighted in connection with specific linguistic structures,

4. This corpus tool is also distributed with the first instalment of the *Corpus of Early English Medical Writing*, the *Middle English Medical Texts* (1375–1500, Taavitsainen et al. 2005).

characteristic for medical texts. Several studies (Jones 1998; Taavitsainen 2001; Pahta & Ratia 2010; Mäkinen 2011) mention efficacy phrases as repetitive – and frequently fixed – elements of medical texts, for instance *uery profitable*, *and it healeth*. Taavitsainen (2001) lists headings and titles as elements of medical texts which tend to remain stable and draws attention to set expressions of Latin origin, discussed also by Marttila (2011). Medical discourse seems to share some fixed phrases with legal discourse, e.g. *the aforesaid/said*+N (Ratia & Suhr 2011) and frequent nominalizations may also be mentioned in this context (see Tyrkkö & Hiltunen 2009). Another recurrent syntactic preference concerns the use of specific prepositional phrases as nominal modifiers, e.g. *in the right side* vs *on the lunge* (Biber et al. 2011). In early medical discourse, there were also typical ways of reporting speech acts and quoting authorities, as revealed by Taavitsainen (2002). Given the range of linguistic structures which fall under the umbrella of "formulaic usage", a fully automatized, corpus-driven extraction method should shed more light on what exactly were the most frequent and most repetitive lexico-syntactic strings. We could then assess the degree of formulaicity in medical texts and establish whether fixed expressions and repetitive constructions constitute an important feature of early medical discourse and, if so, in which particular functions.

3.2 Narrowing down the research questions

With the help of lexical bundles one could, of course, establish the whole inventory of fixed repetitive strings in EMEMT and then look at the structural types and functions of recurrent lexico-syntactic strings, but this inquiry falls outside the scope of the present paper and can be taken up in further research. My intention here is to search for overlaps between the lexical bundles in pre-defined text categories within the corpus and address three major issues:

- To what extent text categories, pre-defined on the basis of their subject matter, share the same lexico-syntactic choices?
- Which bundles occur across text categories?
- Is there a diachronic dimension to the employment of particular lexical bundles?

Thus, the empirical part of this paper will scrutinize the lexical bundles drawn from the normalized version of EMEMT and will concentrate on the bundles which recur in three and more text categories in a given sub-period (see 4.1).

Taavitsainen and Tyrkkö (2010) explain why EMEMT does not employ typological notions such as *genres* or *text types*, but instead it has been chosen

to categorize the texts on extralinguistic grounds – or more specifically, purely on the basis of their subject matter (see also Taavitsainen 2009: 39–40). EMEMT incorporates six text categories, five of which are represented in the full time span covered by the corpus (1500–1700, see Section 4.1). The last category, *The Philosophical Transactions*, covers only the last 35 years. This is a consequence of the fact that the Royal Society was founded in 1660 and the first issue of its journal was published in 1665 (Hiltunen 2010: 127). This paper aims to investigate the shared inventory of lexical bundles in a database which is coherent in terms of time coverage and spans both the sixteenth and the seventeenth centuries. Therefore, the decision has been made to exclude *The Philosophical Transactions* from the present study as they lack the same temporal dimension as the other text categories.[5] A separate synchronic investigation of formulaicity in the late seventeenth century medical discourse would be a welcome complementary project.

The present study is based on the normalized versions (see Sections 2.2–2.3) of the following EMEMT text categories, which together render over 1.6 million words (the acronyms added in brackets will be used in the data presentation in Section 4):

1. General treatises and textbooks (178,416 words, GEN)
2. Treatises on specific topics
   2a. Texts on specific diseases (153,944 words, SPEC)
   2b. Texts on specific methods of diagnosis or treatment (168,098 words, METH)
   2c. Texts on specific therapeutic substances (121,535 words, SUBST)
   2d. Texts on midwifery and children's diseases (102,923 words, CHILD)
   2e. Texts on plague (63,461 words, PLAG)
3. Recipe collections and *materia medica* (338,867 words, REC)
4. Regimens and health guides (208,584 words, REG)
5. Surgical and anatomical treatises (301,701 words, SURG)

The categories range from general to specific and differ in their subject matter. Category 2 is subdivided further to allow finding distinctions between texts on different specific topics. Due to the availability of material, the volume of each text category is different, which will be taken into consideration in data calculations and discussion below. In the analytical sections, I will refer to individual text

5. For a corpus-driven investigation of change and stability in the scientific research article on the basis of the writings of the Royal Society of London, see Atkinson (1999).

categories either by their number (e.g. (1) for the General Treatises), or by their full title, if the context requires it.

When it comes to software settings, the first decision concerned the length of the extracted string. It had to be established what bundle length would render satisfactory results for the research questions posed above. According to the EMEMT compilers, "short clusters of two or three words are very common and usually not particularly insightful" (Tyrkkö, Hickey & Marttila 2010: 256); however, the authors add in a footnote that "…[they] can be of interest too, particularly in corpus-driven analysis of lexical distribution". This study is therefore based on recurrent 3-grams extracted by EMEMT Presenter from the normalized version of the corpus. Research shows that 3-grams are much more numerous than longer bundles (Kopaczyk 2012b: 8), so they supply adequate quantities of analyzable data. Other historical corpus studies, e.g. Culpeper and Kytö (2010) or Kopaczyk (2013), also work with 3-grams, so this choice for medical discourse complements research into other discourse types.

Finally, the cut-off point for drawing lexical bundles from the normalized version of the corpus with the help of EMEMT Presenter was set at five tokens per text. Lower cut-off points have not been used in other studies which employ the lexical bundle methodology, and a higher cut-off point would render fewer lexical bundles to analyze (compare the decisions of individual scholars and the adjustments of the method to specific corpora types and sizes in Kopaczyk 2013: 152–153).

4. Investigating formulaicity in EMEMT text categories

4.1 Degree of overlap: A diachronic outlook

Before proceeding to the analysis proper, it is useful to comment on the proportions in which lexical bundles repeat across text categories depending on the period. It was not clear from the start whether the employment of exactly the same wordings in different text categories would be sustained throughout the corpus span or whether it was limited only to some part of the timeline. I have investigated this issue on the basis of all 3-grams from EMEMT (text categories (1)–(5)), answering two criteria: (a) they had to be present in at least five tokens per text, (b) they had to be present in at least three text categories in a given subperiod. The material has been arranged into six 30-year sub-periods, regardless of text category, and it was determined in how many categories a given bundle appeared during a selected period. Figure 1 shows how many bundles were shared across three or more text categories in EMEMT, depending on the sub-period.

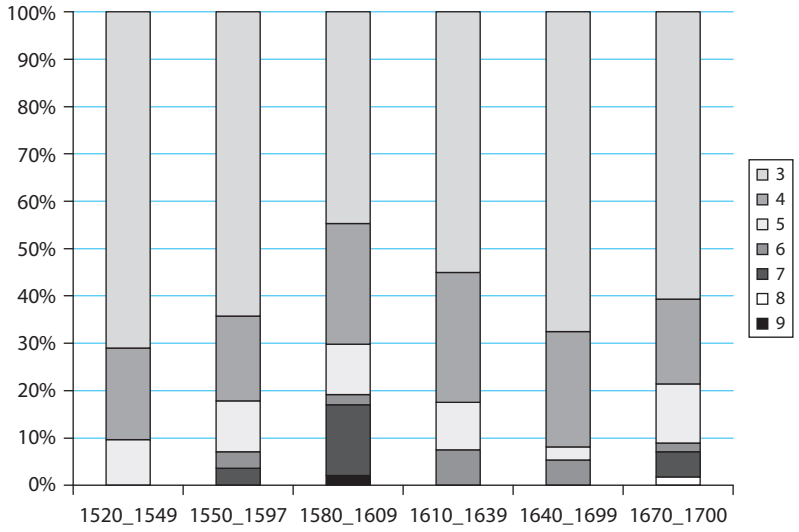

Figure 1. The proportions of shared bundles between three or more text categories in EMEMT, by sub-period

At any given point in time between 1520 and 1700 a relatively large number of shared bundles appeared in three or four text categories. 3-grams shared across the whole corpus, across all the nine text categories specified in 3.2, are very rare. In fact, only one such bundle has been found (see 4.2.2).

It is interesting to note that in the earliest texts the degree of bundle overlap between text categories is the smallest. In other words, in the early sixteenth century the repertoire of fixed lexical strings was less stabilized than in later periods. At the brink of the seventeenth century and one hundred years later, several lexical bundles were shared among seven and more categories of medical texts. It seems that across broadly defined early modern medical discourse, the authors made use of several identical lexico-syntactic patters, regardless of the actual content and topic of a given text. This observation may work as a departure point for a study of textual or discoursal standardization.

The discussion of the overlapping lexical bundles follows in the subsequent sections. As this paper is concerned with the creation of stable patterns of meaning in medical texts, the primary categorization of the extracted bundles has been done on referential semantic grounds (Section 4.2). The bundles which escaped a straightforward semantic categorization were approached from two angles: functional (Section 4.3) and structural (Section 4.4). The functional interpretation was applied to those bundles whose meaning was not purely referential but rather related to their modal and textual properties, e.g. *ought to be* or *according*

to. The final category of structural types includes bundles with no lexical referential or functional content but solely with grammatical content, e.g. *it is a* or *and it is*. In each subsection, the bundles are tabulated by periods and listed alphabetically. Bundle tokens for each category and time span have been counted and the number of occurrences has been normalized (per 10,000 words, separately for each sub-period) to enable cross-categorial comparisons in view of different word counts in categories and sub-periods. This notwithstanding, the discussion below is not meant to be purely quantitative but rather concentrate on the presence or absence of a given bundle in the inventory shared across medical texts on different topics. It is necessary to remember that in order to make it to the tables, a given bundle must have fulfilled two conditions: (1) at least five tokens of the bundle had to be present in a text in a given sub-period; (2) out of these bundles, only those present in at least three text categories were chosen for the analysis. This procedure means that the bundles under scrutiny constitute a shared and repetitive inventory.[6] For easier reference, acronyms for each textual category introduced in Section 3.2, are listed in table headings while category numbers are repeated in the top row of each table, according to the corpus convention.

4.2 Lexical bundle overlaps: Semantic areas

4.2.1 *Quantification, measurements and dosage*

The most prominent semantic area where early modern medical texts display lexical bundle overlaps is quantification. Starting early in the corpus, there is reference to a part or parts of various entities, be it concrete nouns, e.g. body parts (Section 4.2.2) or recipe ingredients (Section 4.2.4), or abstract nouns (Section 4.2.7). We cannot reconstruct the actual modified nouns on the basis of the bundles alone, also because the choice of nouns varies more than the fixed quantification frame. Lexical bundles show that there are certain stable elements in the discourse, which get filled with appropriate diversified content in a given context. The bundles containing reference to a *part* of something continue throughout the periods, overlapping mostly between categories (1), (3) and (5), but also in other types of texts, see Table 2a.

From the mid-seventeenth century onwards, the texts start showing overlaps in specific measurements and dosage, e.g. the Latin *ana +q ii*, or *half an ounce*, shared between (1), (3), and (5).

Interestingly enough, bundles with a specifier *each* start as a common expression for recipes (3) and other types of texts; later they typically surface in texts on midwifery (2d), recipes (3), and surgical texts (5), a trend which continues also in

6. The list of extracted lexical bundles with raw counts for each sub-period is provided in the Appendix.

Table 2a. Lexical bundles expressing quantification, measurements, and dosage: Overlaps across medical genres (1520–1669, 1 – GEN, 2a – SPEC, 2b – METH, 2c – SUBST, 2d – CHILD, 2e – PLAG, 3 – REC, 4 – REG, 5 – SURG; normalized per 10,000 words per sub-period)

	1	2a	2b	2c	2d	2e	3	4	5
P1 1520–1549									
part of the	14.8				2.5		1.1	1.1	16.1
the parts of	2.5		4.9						7.0
P2 1550–1579									
part of the							2.1	7.1	7.3
P3 1580–1609									
as much as	2.0		2.1	1.8					
for the most	1.7					2.0	1.0	1.5	
of each a		1.2		2.9		3.6	1.7		
part of the	3.6	4.5	1.8				1.0		2.9
parts of the	2.3	1.2		4.3	10.8		1.0	1.2	4.8
the most part	2.0					2.6	1.0	1.5	
P4 1610–1639									
ana +q ii	6.4				2.3				2.0
for the most			5.1		2.0		1.5	1.2	
part of the	12.9	3.7						4.2	4.8
parts of the	4.6	4.0			2.0			1.2	1.2
the most part			5.1		2.0		1.7	1.2	
P5 1640–1669									
half an ounce	1.6				10.9	9.0	2.8		1.7
of each half					4.7		0.6		2.2
of each one					2.6		2.4		4.2
of each two					4.1		2.4		1.7
part of the	6.6				4.1		0.8		5.2
parts of the	2.6	3.6			2.6				4.0
the most part	2.6						0.6		1.0
three or four	1.6				3.1		1.8		

the last part of the corpus, with the addition of texts on plague (2e) (see Table 2b). In fact, the final thirty years of the seventeenth century abound in stable, repetitive expressions of quantification, with a larger inventory of specific measurements, e.g. *a spoonful of*, *a pint of*, as well as dosage expressed in the same numbers across different text categories, e.g. *one ounce of*, *two or three*, *three or four*. It turns out

that the most frequent amounts of described or prescribed medicaments were counted in small numbers: between one and four units (see Table 2b).

Table 2b. Lexical bundles expressing quantification, measurements, and dosage: Overlaps across medical genres (1670–1700, 1 – GEN, 2a – SPEC, 2b – METH, 2c – SUBST, 2d – CHILD, 2e – PLAG, 3 – REC, 4 – REG, 5 – SURG; normalized per 10,000 words per sub-period)

	1	2a	2b	2c	2d	2e	3	4	5
P6 1670–1700									
a pint of				2.8			3.5		1.0
a quart of				2.8			0.7	9.5	
a spoonful of	0.8					12.7	1.6		
an ounce of	1.7			3.2		11.2	2.3		0.8
and a half	1.3			1.9		15.9	0.9		
as much as	1.2					17.5			1.8
each half a						9.6	1.0		1.3
each one dram	1.3					9.6			1.3
each two drams	0.8					9.6			0.8
for the most		2.5	1.0		2.4				1.6
half a dram	2.7					25.5	1.2		1.3
half a pint	0.8			1.9			1.6		
half an ounce	4.1			4.1		33.5	3.0		0.8
of each half	1.2					17.5	1.3		1.6
of each one	2.8					23.9	4.2		2.3
of each three	1.0						1.0		0.8
of each two	1.2					11.2	1.9		1.0
one dram of	1.3					8.0	1.2		
one ounce of	2.5						1.2		0.8
ounce of the	1.0			2.8			0.9		
part of the	1.3	2.5	2.9	3.2	5.4		2.8		1.1
parts of the			1.0	2.8			0.9		
quantity of a	0.6						0.7		0.8
the most part		2.5	1.0		2.4				1.6
the quantity of	0.8	4.4		1.6			1.4		1.0
three or four	0.9			3.2		8.0	6.1		
two or three	0.6	2.5		4.4		8.0	4.3	3.4	1.5
two ounces of	0.9					8.0	3.5		1.3

The abundance of measurement types in Period 5 (1640–1669) and especially Period 6 (1670–1700) can be explained by the drive towards greater precision and standardized procedures in medical texts. The core of the lexical bundles is shared between general treatises (1) and other categories, especially surgical texts (5), which may suggest a certain dissemination of fixed linguistic choices from more general to more specific texts. Another potential explanation may be gleaned from authorship investigations which show a preference of specific authors to structure their texts in a typical manner and to employ a recognizable selection of lexico-syntactic combinations (see Tyrkkö, this volume). Some texts in the corpus were written by the same authors but they belong to different text categories, which may have had an impact on the repetitive ways of expressing a particular meaning. Nevertheless, lexical bundles – these containing measurements as well as all the other bundles extracted from the corpus – overlap across all the categories, so this behavior cannot be solely attributed to idiolectal author profiles.

4.2.2 *The body and its parts*

The second major semantic area where medical texts display a range of overlapping lexical bundles is reference to the body and its most important parts, as deemed by early medicine. In view of the normalized counts in Tables 3a–3b, the most repetitive and stable phrase fragment throughout the corpus is *of the body*, which is even shared by all the text categories between 1580 and 1609. This result is not surprising, given that bodily health and malfunction constitute key issues for medicine. In the earliest sixteenth-century texts, the *liver* is the sole part of the body making it into the pool of repetitive shared constructions. Later, while the liver remains in the center of several repetitive bundles, other body parts enter the scene, which can be explained by changes in medical explanations – from the humor theory[7] to explanations based on a better understanding of human anatomy and experiments, for instance to do with blood circulation (Taavitsainen 2010; Taavitsainen et al. 2011). It is also interesting to notice that reference to body parts is shared at first by the more general text categories: treatises (1), recipes (3), regimens (4), and surgical treatises (5), while texts on specific topics (2a–2e) do not seem to share stable patterns in this respect. This distribution changes in the later sub-periods, when especially texts on midwifery (2d) start

7. The liver was perceived as the organ producing one of the humours, the yellow bile, but it was also associated with black bile and blood (Siraisi 1990: 105).

employing structures connected with bodily organs present in the more general medical genres.

Table 3a. Lexical bundles referring to the body and its parts: Overlaps across medical genres (1520–1639, 1 – GEN, 2a – SPEC, 2b – METH, 2c – SUBST, 2d – CHILD, 2e – PLAG, 3 – REC, 4 – REG, 5 – SURG; normalized per 10,000 words per sub-period)

	1	2a	2b	2c	2d	2e	3	4	5
P1 1520–1549									
of man's body	5.9		4.2						12.1
of the body	11.4						1.9	7.8	13.1
of the liver	5.9						4.5	0.9	
the liver and	3.0						3.2	0.9	6.0
P2 1550–1579									
of the body		7.8	1.7				1.7	18.9	6.3
of the head			4.0				2.6		8.3
of the stomach							2.8	2.4	1.0
the body and		2.4						5.9	1.6
the liver and							1.3	2.0	1.6
water of the			1.7			5.0	1.7		
P3 1580–1609									
in our bodies				1.6	5.9		1.0		
of the body	11.6	11.7	7.8	6.1	25.6	5.2	1.0	10.7	9.7
of the head		1.6	2.1				2.1		2.9
of the heart	5.6				4.9	1.6			
the body and	5.6	2.8	1.8					3.2	
the stomach and			2.5	3.6			1.3		
the whole body	2.0		1.8		10.8				3.9
P4 1610–1639									
in the head			6.4		1.7				1.0
of the belly					2.3		1.5		1.2
of the body	11.0	8.3	3.9		1.7			10.4	5.2
of the head	14.7	4.0			3.3				4.0
the whole body	7.3				2.3				1.6

In the second half of the seventeenth century (see Table 3b), there is a growth in the inventory of repetitive structures, while the overlaps continue mostly between the general treatises (1), recipes (3), and surgical texts (5).

Table 3b. Lexical bundles referring to the body and its parts: Overlaps across medical genres (1640–1700, 1 – GEN, 2a – SPEC, 2b – METH, 2c – SUBST, 2d – CHILD, 2e – PLAG, 3 – REC, 4 – REG, 5 – SURG; normalized per 10,000 words per sub-period)

	1	2a	2b	2c	2d	2e	3	4	5
P5 1640–1669									
in the body	2.0	2.3	3.8						1.5
in the head	2.0						1.3		0.8
in the stomach	1.6				3.6			9.9	
of the body	8.5	7.8			4.7		2.2	35.5	11.9
of the head	2.3						0.6		4.4
of the heart	6.6						0.8		17.5
of the liver	3.3						2.8		1.2
of the stomach	2.3				3.1			9.9	2.0
the blood is	2.6	1.6							1.8
the body and	2.6						0.6		1.0
the head and	2.0				2.6		0.9		1.0
the liver and	2.0						2.5		1.2
the stomach and	3.3						0.8	9.9	1.2
the whole body	2.0				2.6				1.3
P6 1670–1700									
in the body	0.8		2.3					6.5	
of the blood		4.4	5.8	5.7			0.7		0.8
of the body	2.6	4.4	3.5	3.5	4.9		1.6	6.1	3.6
the blood and			2.9	1.6		8.0			

The phrase *of the body* is the most repetitive bundle again, but there are new preferences in the texts, e.g. the bundles with reference to *blood*, as one would expect, given the discovery of blood circulation and its impact on medical inquiries of the seventeenth century (Mäkinen 2011: 164). One can also notice that different text categories are "interested" in different body parts. For instance, *heart* appears in shared repetitive fragments of general treatises (1), midwifery texts (2d), recipes (3), and surgical treatises (5), *belly* is of interest in the last three of these, while *head* seems to be scattered across the corpus, with surgical treatises employing fixed reference to *head* in the most consistent manner.

4.2.3 *Time and sequence*

Reference to time and sequence was captured in early medical texts in a stable manner in several phrase fragments, see Table 4.

Table 4. Lexical bundles referring to time and sequence: Overlaps across medical genres (1520–1700, 1 – GEN, 2a – SPEC, 2b – METH, 2c – SUBST, 2d – CHILD, 2e – PLAG, 3 – REC, 4 – REG, 5 – SURG; normalized per 10,000 words per sub-period)

	1	2a	2b	2c	2d	2e	3	4	5
P1 1520–1549									
of the year	2.5		4.9					0.9	
the first is	2.5							1.8	8.0
P2 1550–1579									
the time of		4.4						2.8	0.6
when it is			2.3	4.9			1.3		
P3 1580–1609									
in the first			5.3				1.2	1.2	
in the morning				1.4		2.3	3.1		
when it is	1.7			1.6			1.3		
P4 1610–1639									
in the morning		3.0			1.7		2.7	1.2	2.6
P5 1640–1669									
in the morning	3.6						3.7	11.8	
when it is	2.0						2.3		0.8
P6 1670–1700									
in the beginning	0.9			1.6					1.5
in the morning	5.8			5.7			4.9		1.6
the time of	0.9	2.5	3.5						

In the early sub-periods, fixed time reference is rather general in nature and there is no discernible pattern to text-category preferences. The situation becomes more stable in the later sub-periods, when two specific ingredients of time reference come to the fore in the material. The recurrent beginning of a relative *when*-clause is present across several text categories, with recipes (3) as the main genre employing this sequence marker in a repetitive manner. From Period 3 onwards, the *morning* comes across as an important time when certain procedures need to be carried out or when certain behaviors of the body may be observed (cf. the relatively high scores of this bundle in health regimens (4) in the mid-seventeenth century).

4.2.4 *Ingredients*

Ingredients are typically considered to be a crucial part of recipes (Taavitsainen 2001: 86; Mäkinen 2011: 160). They can therefore be expected to form a conspicuous part of fixed and repetitive expressions in a medical corpus. Bundles containing various ingredients do appear, although they are not as pervasive as the semantic categories discussed above. The frequencies may be quite high, as is the case with *the juice of*, used rather frequently in six text categories in the early seventeenth century (see Table 5), but the inventory of shared bundles is not very numerous.

Table 5. Lexical bundles referring to ingredients: Overlaps across medical genres (1520–1700, 1 – GEN, 2a – SPEC, 2b – METH, 2c – SUBST, 2d – CHILD, 2e – PLAG, 3 – REC, 4 – REG, 5 – SURG; normalized per 10,000 words per sub-period)

	1	2a	2b	2c	2d	2e	3	4	5
P3 1580–1609									
with oil of			2.1	1.6			1.2		2.4
P4 1610–1639									
of an egg		2.0					2.2		1.2
oil of roses		3.0					3.2		2.2
the juice of	6.4	2.3		1.9	2.7	10.8	9.7		
the white of		3.3					4.4		1.2
white of an		2.0					2.2		1.0
P5 1640–1669									
oil of roses					6.2		0.9		1.0
P6 1670–1700									
the juice of	1.8			2.5			6.2		

In fact, shared lexical bundles in this semantic area start emerging only towards the end of the sixteenth century (Period 3). Earlier texts must have made reference to various ingredients of medicaments in a less fixed manner, since the filter for lexical bundles in this study was set at five tokens per text and three overlaps across text categories. In fact, *with oil of* and *oil of roses* were also found in Period 1, but only in midwifery texts (2d). Another pair of syntagmatic overlaps,[8]

8. A *syntagmatic* overlap in lexical bundles can be observed when a bundle of a given length, e.g. a 3-gram, contains elements which constitute the beginning of another 3-gram. Thus, they point to longer repetitive strings. Another type of bundle overlap is a *paradigmatic* overlap, where a shorter bundle is contained in a longer repetitive string. For more discussion and examples, see Kopaczyk (2013: 156–157).

the white of and *of an egg*, were found in Period 1 in midwifery texts (2d) and recipes (3), but they were not frequent enough in other categories. The bundles which did make it to the analysis are therefore quite significant in terms of discourse requirements in a given period. In terms of fixed reference to ingredients, the reference to *oil of roses* or to other potentially useful types of oil (Period 3) surfaces as a stable element of discourse in addition to the juice-related bundles. In seventeenth-century texts, the cluster of lexical bundles referring to *the white of an egg* is shared among texts on specific diseases (2a), recipes (3), and surgical texts (5). This is one of the rare cases among the shared bundles in EMEMT where syntagmatically overlapping 3-grams point towards longer repetitive strings (4- or 5-grams) (see also 4.3.1).

4.2.5 *Quality description*

Several adjectives describing the quality of a given noun have also been found in repetitive strings across the corpus, see Table 6. That noun could typically be a humor or bodily fluid or some aspect of the physical world.

Table 6. Lexical bundles describing quality: Overlaps across medical genres (1520–1700, 1 – GEN, 2a – SPEC, 2b – METH, 2c – SUBST, 2d – CHILD, 2e – PLAG, 3 – REC, 4 – REG, 5 – SURG; normalized per 10,000 words per sub-period)

	1	2a	2b	2c	2d	2e	3	4	5
P1 1520–1549									
hot and dry	3.5						19.7	1.1	
P4 1610–1639									
hot and dry		3.0					2.2	3.7	
P5 1640–1669									
cold and dry	5.6						0.8		2.3
hot and dry	5.9		6.1				4.9		
is hot and	3.0		3.8				1.3		
P6 1670–1700									
cold and moist	1.4		1.2					2.0	
hot and dry	2.3		2.1				1.6		

In terms of types, such bundles are not very numerous, but their presence is quite conspicuous in some text categories, e.g. in general treatises (1), recipes (3), and texts on treatment and diagnosis (2b) from the mid-seventeenth century onwards. Still, bundles describing quality are not shared by more than three categories, and there is a fifty-year gap in the formulaic employment of such expressions in the second half of the sixteenth century. Finally, the important thing to point

out is the fact that all of these formulaic quality descriptions are binomial pairs (or their fragments),[9] a feature which resembles legal discourse (see also 4.3.1).

4.2.6 *Explicit reference to disease and cure*

As disease and cure seem to be potential candidates for syntagmatic overlaps (as in *the cure of the disease*), it was decided to class the relevant bundles together, see Table 7.

Table 7. Lexical bundles referring to disease and cure: Overlaps across medical genres (1520–1700, 1 – GEN, 2a – SPEC, 2b – METH, 2c – SUBST, 2d – CHILD, 2e – PLAG, 3 – REC, 4 – REG, 5 – SURG; normalized per 10,000 words per sub-period)

	1	2a	2b	2c	2d	2e	3	4	5
P2 1550–1579									
of the disease	10.5	2.9				5.0		3.9	
P3 1580–1609									
the cure of	3.0	4.9							4.8
P4 1610–1639									
cure of the	5.5	2.7				9.0			1.4
the cure of	8.3	2.3				9.0			4.8
P6 1670–1700									
the cure of			2.9	2.2		8.0	0.7		1.1

What is interesting is a strong preference for phrasing the reference to cure and disease in the same manner for the most of the corpus time coverage only in four text categories: general treatises (1), texts on specific diseases (2a), texts on plague (2e), and surgical treatises (5). The texts on diagnosis and treatment (2b), for instance, start using the string *the cure of* in a repetitive manner only towards the seventeenth century, and this is also the moment of the most extensive cross-textual overlaps of this bundle.

4.2.7 *Reference to abstract nouns*

The next semantic area where we witness employment of fixed repetitive strings is reference to three abstract nouns: *mind*, *nature*, and *use*, see Table 8.

---

9. There is a substantial body of research on the motivations for binomials (Koskenniemi 1968; Gustafsson 1976), their semantics (Kopaczyk 2009), the arrangement of units in a binomial pair (Cooper & Ross 1975), including the reversibility of element order (Malkiel 1959; Benor & Levy 2006; Mollin 2012). As a discourse feature, these coordinated pairs are typically associated with legal discourse (Mellinkoff 1963; Kopaczyk 2013: 66–71) and other areas where there is a need for mnemonic repetitive devices, all-inclusiveness of reference or a stylistic effect, e.g. alliteration.

Table 8. Lexical bundles referring to abstract nouns: Overlaps across medical genres (1520–1700, 1 – GEN, 2a – SPEC, 2b – METH, 2c – SUBST, 2d – CHILD, 2e – PLAG, 3 – REC, 4 – REG, 5 – SURG; normalized per 10,000 words per sub-period)

	1	2a	2b	2c	2d	2e	3	4	5
P3 1580–1609									
of the mind	3.3	3.7						3.5	
the nature of	4.0	1.9				1.6	1.7	1.2	
P4 1610–1639									
the use of				4.2	3.7			7.4	
P5 1640–1669									
the nature of	3.3	1.6	4.5						
P6 1670–1700									
the use of	2.2		1.9	8.5					

The most prominent bundle seems to be *the nature of*, repeated at different points in time and across the widest variety of texts. It is also crucial to note that the earliest sub-periods in the corpus exhibit no stable lexico-syntactic patterns which would contain abstract reference. This may be attributed either to the lack of textual standardization in medical genres at that time or to the different focus of the earliest medical texts: the bundles drawn from Periods 1 and 2 tend to concern tangible objects such as body parts (see 4.2.2) as well as clarification and efficacy strategies (see 4.3.1 and 4.3.4).

4.2.8 *Reference to humans*

The final semantic category emerging in the lexical bundles in early medical writing is reference to humans. In fact, only mothers inspire a fixed, shared reference across corpus categories, albeit in a very limited temporal and categorial range, compared to other semantic areas, see Table 9.

Table 9. Lexical bundles referring to humans: Overlaps across medical genres (1520–1700, 1 – GEN, 2a – SPEC, 2b – METH, 2c – SUBST, 2d – CHILD, 2e – PLAG, 3 – REC, 4 – REG, 5 – SURG; normalized per 10,000 words per sub-period)

	1	2a	2b	2c	2d	2e	3	4	5
P3 1580–1609									
of the mother			1.8		35.5		1.0		

The prepositional *of*-phrase referring to a mother surfaces in the late-sixteenth century texts on methods of diagnosis and treatment (2b) and in recipes (3), but it is clearly most frequent and repetitive in midwifery texts (2d).

4.3 Lexical bundle overlaps: Functional areas

4.3.1 *Clarification*

The most prominent function of shared lexical bundles in early medical discourse, apart from the referential contexts discussed above, is clarification and explanation. This large functional category contains strings to do with cause and effect (e.g. *by reason of*), fixed phrases introducing reference to authority (e.g. *according to the*), relative clause fragments with *which* and *that*, and cohesion markers making reference to earlier discourse (e.g. *of the said*). All these linguistic tools help to clarify the contents of the text, make the text more explicit, establish links between parts of the text for better understanding, and support its message with intertextual references. The corpus does not reveal any preference for overlaps between specific text categories in this respect. Clarification bundles are scattered across all categories, although in selected sub-periods some categories may have more in common, for instance general treatises (1), regimens (4), and surgical treatises (5) in Period 1; recipes (3), and surgical treatises (5) in Period 2; general treatises (1), texts on specific diseases (2a), and texts on plague (2e) in Period 3; and texts on therapeutic substances (2c), regimens (4), and surgical treatises (5) in Period 4, as specified in Table 10.

Table 10. Lexical bundles expressing clarification: Overlaps across medical genres (1520–1700, 1 – GEN, 2a – SPEC, 2b – METH, 2c – SUBST, 2d – CHILD, 2e – PLAG, 3 – REC, 4 – REG, 5 – SURG; normalized per 10,000 words per sub-period)

	1	2a	2b	2c	2d	2e	3	4	5
P1 1520–1549									
by reason of			3.5		4.6			1.1	
is to say	4.9		7.6					1.2	5.0
it is called	2.5						5.8	1.9	
of the same	12.3				6.1				15.1
of the which	7.9				5.1			3.4	
that is to	4.4		7.6					5.3	7.0
therefore it is	3.5						1.3		5.0
P2 1550–1579									
according to the		2.4	2.0	8.8				2.8	0.6
by reason of		2.4					3.0	4.3	
is to be				6.9				4.7	5.2
is to say				29.4		5.9	1.3		7.8
it is called		3.4					1.1		1.4
of the same		10.2	3.7				1.7		2.6
that is to				30.4		5.9	1.3		10.3

(*Continued*)

Table 10. (Continued)

	1	2a	2b	2c	2d	2e	3	4	5
that which is	6.6					5.0			1.5
P3 1580–1609									
according to the	3.0	1.2	1.8		12.8	2.3		1.7	
are to be		1.2		1.1		1.6			
as it is	4.0	1.9					1.2		3.4
as it were	5.0	8.4					2.3		
by reason of		1.2		2.5	6.9	2.0		2.2	
in respect of		1.2		1.6			1.2		
is to say	3.3					2.3		1.7	
is to be	7.3	2.8	2.1	4.1			1.0		
of the said		1.6					1.0		4.8
of the same	3.0	4.9	2.8	1.1		5.5	1.5		7.7
that is to	8.3	1.4				2.6		1.7	
that it is	5.3			6.5		1.6	2.7	2.7	
that which is	6.6	1.6							4.4
P4 1610–1639									
according to the				1.9	1.7			1.2	5.0
because it is		2.0		1.9				1.7	1.2
by reason of	9.2	4.3		3.8	3.7			6.7	2.8
is to be		6.0						1.7	1.0
of the same		2.7		4.2					1.0
reason of the				1.9				3.5	1.8
that which is		2.3						3.2	2.6
P5 1640–1669									
according to the	1.6		6.8				1.0	9.9	
by reason of	2.0	2.3	3.8						4.7
is to be			3.8				2.2		4.0
P6 1670–1700									
according to the	2.8	2.5	1.5	3.8			1.7	1.7	2.8
are to be		1.2		1.1		1.6			
in the same				1.6		8.0	1.2		
is to be	7.3	2.8	2.1	4.1			1.0		
it is to				1.6			0.7		1.3
the cause of	0.9		1.2	1.6					
those that are	2.1					8.0	2.2		
which is the				1.6			0.7	2.0	

What is striking in comparison to other types of lexical bundles is that clarification bundles are present throughout the corpus in more or less the same inventory and with a relatively large degree of individual cross-textual overlaps. For instance, the bundle *according to the* appears in four to seven textual categories from the mid-sixteenth century till 1700. It seems that authors of texts on various medical topics chose to make links to other texts and authorities. This preference for referring to authority in a fixed manner comes across a consequence of the preoccupation of medicine with the transmission of knowledge (Taavitsainen 2010). Still, the diachronic results are unexpected in view of earlier research, which has established that references to authority and shared community knowledge diminished in time (Hiltunen & Tyrkkö 2011: 72; Marttila 2011: 148–151).

Several clarification bundles resemble formulaic choices present in legal texts. A recent study on early Scots legal and administrative texts revealed that cohesion-related bundles such as *of the same*, *of the said*, *that is to say* constitute the core of textual fixedness in early legal discourse (Kopaczyk 2013).[10] Previous research on early English medical texts has touched upon points of contact between these two registers: the language of medicine and the language of the law. In Ratia and Suhr's (2011) study on medical pamphlets, the recurrent deictic constructions resemble those found in the legal texts from the *Lampeter Corpus* by Claridge (2001) and in Scots legal discourse by Kopaczyk (2013). The inventory of recurrent lexical strings which are shared across different medical text categories includes *(that) is to say* / *is to be* (potential 4-grams, indicated by syntagmatic overlaps in all sub-periods), *of the said*, *of the same*, *of the which*, and *in the same*. In the explanatory domain, the appearance of the noun *cause* in the final years of the seventeenth century may serve as an indication of a more frequent discussion of causes and effects revealed by experimental practices around that time.

4.3.2 *Conditionals*

Medical texts also contain formulaic elements of conditional structures. The inventory of these bundles is limited (see Table 11), but the string *if it be* was employed by authors across all sub-periods with quite a large degree of text category overlaps. In the mid-sixteenth century (Period 2), for instance, this bundle appears throughout a spectrum of seven textual categories of medical discourse and continues to be used in later sub-periods.

---

10. As a genetically related language, Scots may exhibit similar structural preferences to contemporary English, especially in view of a significant degree of contact between the two languages. Scots legal discourse, however, may exhibit its own unique characteristics due to the historical and cultural background of the Scots law and the extralinguistic conditions in which early legal texts were compiled (Kopaczyk 2013). Unfortunately, corpus-driven studies on textual fixedness in historical English legal discourse have not been attempted yet.

Table 11. Lexical bundles expressing condition: Overlaps across medical genres (1520–1700, 1 – GEN, 2a – SPEC, 2b – METH, 2c – SUBST, 2d – CHILD, 2e – PLAG, 3 – REC, 4 – REG, 5 – SURG; normalized per 10,000 words per sub-period)

	1	2a	2b	2c	2d	2e	3	4	5
P1 1520–1549									
if a man	2.5						6.9	0.9	
if it be	3.9		12.5		2.5		15.0	4.4	
if they be	2.5						1.5	3.7	
P2 1550–1579									
if a man			2.7			5.0	2.1		
if it be	14.4		5.0	17.6		5.0	1.7	5.1	3.2
P3 1580–1609									
if it be			5.6	1.1				2.5	
P4 1610–1639									
if it be		2.0		4.2	8.0		11.9	1.7	1.2
whether it be			3.2	2.3	2.7				
P5 1640–1669									
if it be	1.6	2.9			2.6		2.7	19.7	4.2
P6 1670–1700									
if it be	2.8						1.0		2.1

The formulaic behavior of conditional phrase fragments points to the fact that in many medical texts there was a discussion or exposition of open-ended situations. The actions described depended on the developments which could but may not have occurred, and the authors needed to make reference to potential scenarios. It is intriguing that they should do this by means of a relatively stable lexico-syntactic frame throughout the corpus.

4.3.3 *Modality and hedges*

Predictions, prognostications, and recommendations are expressed in medical discourse by various stance markers.[11] The phrase fragments surfacing in lexical

11. Biber's (2004) study shows that medical discourse scores relatively low in terms of stance markers, compared to other contemporary genres (1650–1990). There are, however, certain preferences for expressing stance by means of specific linguistic tools in medical texts, e.g. stance adverbials. Gray et al. (2011) follow up this research, concentrating on the expressions of stance in the last part of EMEMT, with focus on the *Philosophical Transactions*.

bundles include modal verbs such as *may*, *shall* and *ought to*, which convey epistemic modality (what may happen) and deontic modality (what should happen). In earlier texts, predictions with *it may be* behave in a formulaic manner across medical genres, see Table 12. Advice or instruction (see also 4.3.5) is centered around repetitive fragments with *ought to be* and *it shall be*.

Table 12. Lexical bundles expressing modality and hedges: Overlaps across medical genres (1520–1700, 1 – GEN, 2a – SPEC, 2b – METH, 2c – SUBST, 2d – CHILD, 2e – PLAG, 3 – REC, 4 – REG, 5 – SURG; normalized per 10,000 words per sub-period)

	1	**2a**	**2b**	**2c**	**2d**	**2e**	**3**	**4**	**5**
P1 1520–1549									
it may be	3.0				5.1		18.2		
P2 1550–1579									
ought to be			2.7	14.7			1.7	5.1	1.4
P3 1580–1609									
and such like	1.7					2.3	1.0	3.0	
it may be	2.3	1.2		1.6		1.6			
P4 1610–1639									
a kind of				1.9	1.7			1.7	
and such like		2.3		1.9			1.5		2.2
as it were				3.8	2.3				1.8
it may be				1.9	2.0			1.2	1.8
it shall be		2.7					1.7		1.4
P5 1640–1669									
as it were	3.3						0.6		2.7
P6 1670–1700									
all sorts of	0.8						1.0		1.1
and such like			2.5				0.9	2.7	
ought to be	1.3		2.5				1.4	1.7	1.0

In predictions one tries to safeguard his or her claims against a potential lack of success. This may be one of the reasons why hedges appear in medical discourse (Atkinson 1999; Gotti 2011:211–215),[12] and they do so in a formulaic format

12. Both Atkinson and Gotti study the language of the *Philosophical Transactions*. Their explanations for the employment of stance expressions by seventeenth-century scientists should be extrapolated to earlier medical texts with caution.

across different text categories. Hedging bundles in EMEMT include *and such like*, *a kind of*, *all sorts of* and *as it were*. They are shared by texts on therapeutic substances (2c) in the earlier part of the corpus and in surgical treatises (5) in the latter part. The employment of recurrent hedges and modal expressions is most consistent in recipes (3) in all the sub-periods.

4.3.4 *Efficacy phrase fragments*

It is intriguing that efficacy phrases should behave in a formulaic manner only up to the first half of the seventeenth century (see Table 13). Lexical bundles reveal several lexico-syntactic arrangements which are shared across text categories in EMEMT; all of them, nevertheless, contain the adjective *good*. Mäkinen's (2011) research[13] describes various means of expressing assurance that a given medical solution is indeed worth following. It seems, however, that only the strings with the adjective *good* were fixed enough to make it to the formulaic pool of shared expressions.

Table 13. Lexical bundles containing efficacy phrases: Overlaps across medical genres (1520–1700, 1 – GEN, 2a – SPEC, 2b – METH, 2c – SUBST, 2d – CHILD, 2e – PLAG, 3 – REC, 4 – REG, 5 – SURG; normalized per 10,000 words per sub-period)

	1	2a	2b	2c	2d	2e	3	4	5
P1 1520–1549									
good for the					3.0		6.4	1.6	
is good to			11.8		6.1		14.4	1.1	
it is good			7.6		6.6		16.9	1.6	
P2 1550–1579									
is good to			16.3			5.0	4.3		
it is good			16.7			13.9	5.8		
P3 1580–1609									
it is good			6.0				2.5	1.2	
P4 1610–1639									
is very good							1.2	2.0	1.4

When it comes to text categories which display a preference for fixed efficacy strings, there is no surprise that texts on diagnostic methods (2b) should aim to persuade the reader that the methods are efficient. Similarly, recipes also feature

13. That investigation was carried out on the basis of Category 3, recipes and *materia medica* from EMEMT.

such lexical bundles. In fact, the authors of the EMEMT Presenter manual have discovered that the string *it is good for* is the most formulaic elements in recipes in their illustrative sample search (Tyrkkö, Hickey & Marttila 2010: 259–260). To a lesser extent, recurrent efficacy phrases are found in texts on midwifery (in Period 1) and health regimens (in Period 1 and 3–4).

4.3.5 *Directives*

It is intriguing to find so few recurrent directive patterns in medical texts (see Table 14), even though medical writings are often concerned with instructing a practitioner or giving advice and guidance. Studies have shown that elements of interaction are an important part of medical discourse (also depending on the audience, see Marttila 2011) but it transpires that this function is not formulaic enough to make it to the pool of recurrent shared bundles.

Table 14. Lexical bundles expressing directives: Overlaps across medical genres (1520–1700, 1 – GEN, 2a – SPEC, 2b – METH, 2c – SUBST, 2d – CHILD, 2e – PLAG, 3 – REC, 4 – REG, 5 – SURG; normalized per 10,000 words per sub-period)

	1	2a	2b	2c	2d	2e	3	4	5
P1 1520–1549									
lay it to					2.5		5.8	0.9	
P4 1610–1639									
let the patient	4.6						3.7		1.0
P6 1670–1700									
take of the	5.9					12.7	3.5		

Due to the structural character of the findings, we are not informed about the right-hand and left-hand co-text in which a particular bundle occurred. Manual checks reveal that the recurrent directives were part of longer instructions, as in examples (2a–c), and could also appear in modal passages (as in (2b–c); cf. Section 4.3.3):

(2) a. Take Mastick, Frankincense, …, make them into powder, and with the juice of Mints … make a Pultis, and *lay it to* the stomach.
<1653, Pemell, *De Morbis Puerorum*>

b. Take and stampe it & fry it with shepes talow/ and make a playster/ and *lay it to* a potager man/ & it shall helpe hym within.iii. dayes
<1525, *Newe Matter*>

c. Put oyle of Hempseede warme into the eare, and stop it with sheeps wooll, and *let the Patient* leape and use exercise, then lye downe on the side that he is payned, to see if any thing will run out.
<1634, Hawes, *Pooremans Plasterbox*>

The bundles extracted from the corpus highlight recipes (3) as the genre which shares directives with other text categories. The modest inventory of shared bundles prevents more general conclusions though.

4.4 Lexical bundle overlaps: Structural frames

4.4.1 *Copula verbs*

As a corpus-driven method of automatic extraction, lexical bundles often render chunks of discourse which are not easily classifiable on semantic or discoursal grounds. This is why it seems appropriate to have a separate category for structural frames focusing on a recurrent grammatical element. Table 15 presents lexical bundles drawn from EMEMT, which center around the copula verb *to be*, surfacing, in fact, only in its finite form in the third person singular.

Table 15. Lexical bundles centered around a copula verb: Overlaps across medical genres (1520–1700, 1 – GEN, 2a – SPEC, 2b – METH, 2c – SUBST, 2d – CHILD, 2e – PLAG, 3 – REC, 4 – REG, 5 – SURG; normalized per 10,000 words per sub-period)

	1	2a	2b	2c	2d	2e	3	4	5
P1 1520–1549									
and it is							1.7	2.0	5.9
it is a	11.8					5.9	1.1		1.6
it is not				6.9		5.0			0.7
P2 1550–1579									
and it is							1.7	2.0	5.9
it is a	11.8					5.9	1.1		1.6
it is not				6.9		5.0			0.7
P3 1580–1609									
it is a			1.8	1.1		2.6	1.9		
P4 1610–1639									
it is a			7.1	3.1	1.7		1.7		1.4
that it is		2.7	7.7	1.9			2.2		
there is a			3.2	3.1	1.7				
P5 1640–1669									
it is a		2.3					0.6		2.9
P6 1670–1700									
it is a			1.0		2.4		6.5		

These bundles may indicate syntagmatic overlaps with other bundles, but they may also function as fixed grammatical hubs, allowing for a variety of complements and modifiers to appear on both sides. Fragments such as *it is a* tend to remain stable and recurrent and only serve to introduce changeable content, as appropriate in a given context. Practically all text categories share a preference for this particular bundle at some point in the corpus timeline.

4.4.2 *Prepositional phrase fragments*

The final category of lexical bundles in EMEMT gathers fragments of prepositional phrases (Table 16). As in the previous section, these bundles have no particular referential or discoursal function but rather they provide a frame for noun phrases, typically indicated by the bundles finishing in *the*, to complement these stable fragments and bring in referential meaning.

Table 16. Lexical bundles centered on a preposition: Overlaps across medical genres (1520–1700, 1 – GEN, 2a – SPEC, 2b – METH, 2c – SUBST, 2d – CHILD, 2e – PLAG, 3 – REC, 4 – REG, 5 – SURG; normalized per 10,000 words per sub-period)

	1	2a	2b	2c	2d	2e	3	4	5
P1 1520–1549									
and in the			6.3				2.1		17.1
in to the							1.3	0.9	6.0
it to the					6.1		4.1	0.9	
P2 1550–1579									
and of the				4.9			3.0		2.1
out of the		2.4	6.7			7.9	4.9	2.8	0.6
P3 1580–1609									
and in the				1.4			1.0	1.7	
out of the	2.3	1.2	4.9	1.6		2.9	1.3		9.7
P5 1640–1669									
and in the	2.3						1.8		2.9
out of the		2.3			2.6		1.1		6.0

The most popular prepositional phrase fragments across the text categories and sub-periods are *and in the*, which refers to position or direction, and *out of the*, which typically indicates a direction of movement, or a choice from a range of options. Both are especially popular in recipes (3) and surgical treatises (1), with a notable absence of such strings in general treatises, at least in a recurrent format.

5. General observations and further research

The material rendered by automatic lexical bundle extraction is very diverse in terms of form, function, and token numbers, which may make it difficult to analyze. The introduction of an additional parameter, the textual overlap, has efficiently reduced the number of 3-grams drawn from the files with the help of EMEMT Presenter, leaving only those which constituted the unchangeable lexico-syntactic core of the corpus. Thus, the inventory of lexical bundles occurring across at least three text categories has been established.

The recurrent lexical bundles have been divided into three major groups: those with a referential or semantic motivation (Section 4.2), those whose presence is connected with non-referential meaning and with a specific discourse function (Section 4.3), and those which create stable grammatical frames for exchangeable content (Section 4.4). In each of these groups, I introduced further subdivisions where individual lexical bundles were traced across the text categories and sub-periods. The analysis revealed affinities between types of medical texts, pre-defined on the basis of extralinguistic criteria in the corpus. It is the general treatises (1) that share the most lexical bundles with other text categories throughout the corpus. In the semantic areas, the fixed lexico-syntactic choices in general treatises (1) most often overlap with surgical treatises (5), recipe collections (3), and midwifery and children's diseases (2d), especially in reference to body parts and quantification. In the functional areas, the overlaps among the text categories depend to a large extent on the period but, for example, texts on specific diseases (2a), specific therapeutic substances (2c), and on plague (2e) share more bundles here than in the purely referential domain. Finally, the most typical grammatical frames are to be found mostly across recipe collections (3) and surgical treatises (5).

When it comes to the diachronic dimension, the analysis has shown that different lexical bundles gain prominence in different periods. Efficacy phrase fragments are often shared across text categories between 1520–1639; similarly, clarification bundles dwindle after 1640; ingredients are captured in several strings repeated most frequently at the beginning of the seventeenth century; body parts are the focus of lexical bundles between 1640–1669; all kinds of measurements become prominent between 1640–1700. These findings should complement earlier research with more specific conclusions on the formulaic behavior of the linguistic structures under study.

Apart from the positive conclusions, for instance the support for formulaicity of recipes as discussed by Marttila (2010: 103–104), one may also draw negative conclusions. The categories of texts which share the least common patterns are specific treatises and especially texts on plague (2e) and therapeutic substances

(2c). This may be due to the fact that these areas of early medicine were quite specialized in their object of interest and, therefore, there were few opportunities to phrase the message in the same way as other contemporary medical texts. It would be interesting to follow up the present inquiry with an assessment of formulaicity in individual text categories in EMEMT or concentrate on the most frequent lexical bundles to shed light on the areas of medical discourse which were most prone to textual fixedness.

Data

EMEMT = *Early Modern English Medical Texts*. 2010. Compiled by Irma Taavitsainen, Päivi Pahta, Turo Hiltunen, Martti Mäkinen, Ville Marttila, Maura Ratia, Carla Suhr, and Jukka Tyrkkö. Corpus Presenter software by Raymond Hickey. Amsterdam: John Benjamins.

References

Altenberg, Bengt. 1998. On the phraseology of spoken English: The evidence of recurrent word-combinations. In *Phraseology: Theory, Analysis, and Applications*, Anthony Paul Cowie (ed.), 101–122. Oxford: Clarendon Press.

Ari, Omer. 2006. Review of three software programs designed to identify lexical bundles. *Language Learning and Technology* 10(1): 30–37.

Atkinson, Dwight. 1999. *Scientific Discourse in Sociohistorical Context: The Philosophical Transactions of the Royal Society of London, 1675–1975*. Mahwah NJ: Lawrence Erlbaum Associates.

Baron, Alistair. 2010. VARD 2. ⟨http://www.comp.lancs.ac.uk/~barona/vard2⟩ (December 2012).

Baron, Alistair & Rayson, Paul. 2008. VARD 2: A tool for dealing with spelling variation in historical corpora. *Proceedings of the Postgraduate Conference in Corpus Linguistics, Aston University, Birmingham, UK, May 2008*. ⟨http://acorn.aston.ac.uk/conf_proceedings.html⟩ (December 2012).

Benor, Sarah Bunin & Levy, Roger. 2006. The chicken or the egg? A probabilistic analysis of English binomials. *Language* 82(2): 233–278.

Biber, Douglas. 2004. Historical patterns for the grammatical marking of stance: A cross-register comparison. *Journal of Historical Pragmatics* 5(1): 107–136.

Biber, Douglas. 2009. A corpus-driven approach to formulaic language in English: Multi-word patterns in speech and writing. *International Journal of Corpus Linguistics* 14(3): 275–311.

Biber, Douglas & Barbieri, Federica. 2007. Lexical bundles in university spoken and written registers. *English for Specific Purposes* 26: 263–286.

Biber, Douglas, Johansson, Stig, Leech, Geoffrey N., Conrad, Susan & Finegan, Edward. 1999. *Longman Grammar of Spoken and Written English*. London: Longman.

Biber, Douglas, Conrad, Susan & Cortes, Viviana. 2004. If you look at… Lexical bundles in university teaching and textbooks. *Applied Linguistics* 25(3): 371–405.

Biber, Douglas, Gray, Bethany, Honkapohja, Alpo & Pahta, Päivi. 2011. Prepositional modifiers in Early English medical prose. A study on their historical development in noun phrases. In *Communicating Early English Manuscripts*, Päivi Pahta & Andreas H. Jucker (eds), 197–211. Cambridge: CUP.

Blake, Norman F. & Robinson, Peter M.W. (eds). 1993. *The Canterbury Tales Project Occasional Papers,* Vol. 1. Oxford: Office for Humanities Communication.

Claridge, Claudia. 2001. Structuring text: Discourse deixis in Early Modern English texts. *Journal of English Linguistics* 29(1): 55–71.

Conklin, Kathy & Schmitt, Norbert. 2008. Formulaic sequences: Are they processed more quickly than nonformulaic language by native and nonnative speakers? *Applied Linguistics* 29(1): 72–89.

Cooper, William E. & Ross, John R. 1975. World order. In *Papers from the Parasession on Functionalism*, Robin E. Grossman, L. James San & Timothy J. Vance (eds), 63–111. Chicago IL: Chicago Linguistic Society.

Culpeper, Jonathan & Kytö, Merja. 2002. Lexical bundles in Early Modern English: A window into the speech-related language of the past. In *Sounds, Words, Texts, and Change: Selected Papers from 11 ICEHL* [Current Issues in Linguistic Theory 224], Teresa Fanego, Belén Méndez-Naya & Elena Seoane (eds), 45–63. Amsterdam: John Benjamins.

Culpeper, Jonathan & Kytö, Merja. 2010. *Early Modern English Dialogues: Spoken Interaction as Writing*. Cambridge: CUP.

Danet, Brenda. 1980. Language in the legal process. *Law and Society Review* 14(3): 445–564.

Gotti, Maurizio. 2011. The development of specialized discourse in the *Philosophical Transactions*. In *Medical Writing in Early Modern English*, Irma Taavitsainen & Päivi Pahta (eds), 204–220. Cambridge: CUP.

Gray, Bethany, Biber, Douglas & Hiltunen, Turo. 2011. The expression of stance in early (1665–1712) publications of the *Philosophical Transactions* and other contemporary medical prose: Innovations in a pioneering discourse. In *Medical Writing in Early Modern English*, Irma Taavitsainen & Päivi Pahta (eds), 221–247. Cambridge: CUP.

Gustafsson, Marita. 1976. The frequency and 'frozenness' of some English binomials. *Neuphilologische Mitteilungen* 77: 623–637.

Halliday, Michael A.K. 1978. *Language as a Social Semiotic. The Social Interpretation of Language and Meaning*. London: Edward Arnold.

Hiltunen, Turo. 2010. *Philosophical Transactions*. In *Early Modern English Medical Texts. Corpus Description and Studies*, Irma Taavitsainen & Päivi Pahta (eds), 127–131. Amsterdam: John Benjamins.

Hiltunen, Turo & Tyrkkö, Jukka. 2011. Verbs of knowing: Discursive practices in Early Modern vernacular medicine. In *Medical Writing in Early Modern English*, Irma Taavitsainen & Päivi Pahta (eds), 44–73. Cambridge: CUP.

Jones, Claire. 1998. Formula and formulation: 'Efficacy phrases' in medieval English medical manuscripts. *Neuphilologische Mitteilungen* 99(2): 199–209.

Kohnen, Thomas. 2010. Religious discourse. In *Historical Pragmatics*, Andreas H. Jucker & Irma Taavitsainen (eds), 523–547. Berlin and New York: De Gruyter Mouton.

Kopaczyk, Joanna. 2009. Multi-word units of meaning in 16th-century legal Scots. In *Selected Proceedings of the 2008 Symposium on New Approaches in English Historical Lexis (HELLEX 2)*, Rod W. McConchie, Alpo Honkapohja & Jukka Tyrkkö (eds), 88–95. Somerville, MA: Cascadilla Proceedings Project.

Kopaczyk, Joanna. 2012a. Applications of the lexical bundles method in historical corpus research. In *Corpus Data Across Languages and Disciplines*, Piotr Pęzik (ed.), 83–95. Frankfurt a/Main: Peter Lang.

Kopaczyk, Joanna. 2012b. Long lexical bundles and standardisation in historical legal texts. *Studia Anglica Posnaniensia* 47(2–3): 3–25.

Kopaczyk, Joanna. 2013. *The Legal Language of Scottish Burghs. Standardization and Lexical Bundles, 1380–1560*. Oxford and New York: OUP.

Koskenniemi, Inna. 1968. *Repetitive Word-pairs in Old and Early Middle English Prose*. Turku: Turun Yliopisto.

Kytö, Merja. 2012. New perspectives, theories, and methods: Corpus linguistics. In *English Historical Linguistics. An International Handbook*, Vol. 2, Alexander Bergs & Laurel J. Brinton (eds), 1509–1531. Berlin and Boston: De Gruyter Mouton.

Kytö, Merja & Smitterberg, Erik. 2006. 19th-century English: An age of stability or a period of change? In *Corpus-based Studies of Diachronic English*, Roberta Facchinetti & Matti Rissanen (eds), 199–230. Bern: Peter Lang.

Lehto, Anu, Baron, Alistair, Ratia, Maura & Rayson, Paul. 2010. Improving the precision of corpus methods. In *Early Modern English Medical Texts. Corpus Description and Studies*, Irma Taavitsainen & Päivi Pahta (eds), 279–289. Amsterdam: John Benjamins.

Mäkinen, Martti. 2011. Efficacy phrases in Early Modern English medical recipes. In *Medical Writing in Early Modern English*, Irma Taavitsainen & Päivi Pahta (eds), 158–179. Cambridge: CUP.

Malkiel, Yakov. 1959. Studies in irreversible binomials. *Lingua* 8: 113–160.

Marttila, Ville. 2010. Category 3. Recipe collections and *materia medica*. In *Early Modern English Medical Texts. Corpus Description and Studies*, Irma Taavitsainen & Päivi Pahta (eds), 101–109. Amsterdam: John Benjamins.

Marttila, Ville. 2011. New arguments for new audiences: A corpus-based analysis of interpersonal strategies in Early Modern English medical recipes. In *Medical Writing in Early Modern English*, Irma Taavitsainen & Päivi Pahta (eds), 135–157. Cambridge: CUP.

Mellinkoff, David. 1963. *The Language of the Law*. Boston MA: Little Brown.

Mollin, Sandra. 2012. Revisiting binomial order in English: Ordering constraints and reversibility. *English Language and Linguistics* 16(1): 83–103.

Pahta, Päivi & Ratia, Maura. 2010. Treatises on specific topics. In *Early Modern English Medical Texts. Corpus Description and Studies*, Irma Taavitsainen & Päivi Pahta (eds), 73–99. Amsterdam: John Benjamins.

Ratia, Maura & Suhr, Carla. 2011. Medical pamphlets: Controversy and advertising. In *Medical Writing in Early Modern English*, Irma Taavitsainen & Päivi Pahta (eds), 180–203. Cambridge: CUP.

Rayson, Paul, Archer, Dawn, Baron, Alistair & Smith, Nicholas. 2007. Tagging historical corpora – The problem of spelling variation. In *Proceedings of Digital Historical Corpora, Dagstuhl-seminar 06491, Wadern, Germany, 3–8 December 2006.* ISSN 1862–4405.

Robinson, Peter. 2009. What text is really not, and why editors have to learn to swim. *Literary and Linguistic Computing* 24(1): 41–52.

Scott, Michael. 1997. PC analysis of key words – and why key key words. *System* 25(1): 1–13.

Sinclair, John. 1991. *Corpus, Concordance and Collocation.* Oxford: OUP.

Siraisi, Nancy G. 1990. *Medieval and Early Renaissance Medicine. An Introduction to Knowledge and Practice*. Chicago IL: The University of Chicago Press.

Stubbs, Michael & Barth, Isabel. 2003. Using recurrent phrases and text-type discriminators: A quantitative method and some findings. *Functions of Language* 10(1): 61–64.

Taavitsainen, Irma. 2001. Middle English recipes. Genre characteristics, text type features and underlying traditions of writing. *Journal of Historical Pragmatics* 2(1): 85–113.

Taavitsainen, Irma. 2002. Historical discourse analysis. Scientific language and changing thought-styles. In *Sounds, Words, Texts, and Change. Selected Papers from 11 ICEHL* [Current Issues in Linguistic Theory 224], Teresa Fanego, Belén Méndez-Naya & Elena Seoane (eds), 201–226. Amsterdam: John Benjamins.

Taavitsainen, Irma. 2009. The pragmatics of knowledge and meaning: Corpus linguistic approaches to changing thought-styles in Early Modern English medical discourse. In *Corpora: Pragmatics and Discourse. Papers from the 29th International Conference on English Language Research on Computerized Corpora* (ICAME 29)*, Ascona, Switzerland, 14–18 May 2008*, Andreas H. Jucker, Daniel Schreier & Marianne Hundt (eds), 37–62. Amsterdam: Rodopi.

Taavitsainen, Irma. 2010. Discourse and genre dynamics in Early Modern English medical writing. In *Early Modern English Medical Texts. Corpus Description and Studies*, Irma Taavitsainen & Päivi Pahta (eds), 29–53. Amsterdam: John Benjamins.

Taavitsainen, Irma, Pahta, Päivi & Mäkinen, Martti (eds). 2005. *Middle English Medical Texts*. Amsterdam: John Benjamins.

Taavitsainen, Irma & Pahta, Päivi (eds). 2010. *Early Modern English Medical Texts. Corpus Description and Studies*. Amsterdam: John Benjamins.

Taavitsainen, Irma & Tyrkkö, Jukka. 2010. The field of medical writing with fuzzy edges. In *Early Modern English Medical Texts. Corpus Description and Studies*, Irma Taavitsainen & Päivi Pahta (eds), 57–61. Amsterdam: John Benjamins.

Taavitsainen, Irma, Jones, Peter Murray, Pahta, Päivi, Hiltunen, Turo, Marttila, Ville, Ratia, Maura, Suhr, Carla & Tyrkkö, Jukka. 2011. Medical texts in 1500–1700 and the *Corpus of Early Modern English Medical Texts*. In *Medical Writing in Early Modern English*, Irma Taavitsainen & Päivi Pahta (eds), 9–25. Cambridge: CUP.

Tiersma, Peter M. 1999. *Legal Language*. Chicago IL: University of Chicago Press.

Tiersma, Peter M. 2006. Some myths about legal language. *Law, Culture and the Humanities* 2: 29–50.

Tognini-Bonelli, Elena. 2001. *Corpus Linguistics at Work* [Studies in Corpus Linguistics 6]. Amsterdam: John Benjamins.

Tyrkkö, Jukka & Hiltunen, Turo. 2009. Frequency of nominalization in Early Modern English medical writing. In *Corpora: Pragmatics and Discourse. Papers from the 29th International Conference on English Language Research on Computerized Corpora*, Andreas Jucker, Marianne Hundt & Daniel Schreier (eds), 293–316. Amsterdam: Rodopi.

Tyrkkö, Jukka, Hickey, Raymond & Marttila, Ville. 2010. Exploring Early Modern English medical texts. In *Early Modern English Medical Texts. Corpus Description and Studies*, Irma Taavitsainen & Päivi Pahta (eds), 219–277. Amsterdam: John Benjamins.

Appendix

Raw counts of lexical bundles extracted from EMEMT categories (1)–(5) by sub-period.

P1 1520–1549	1	2a	2b	2c	2d	2e	3	4	5
and in the			9				10		17
and it is	13						59		16
by reason of			5		9			6	
good for the					6		30	9	
hot and dry	7						92	6	
if a man	5						32	5	
if it be	8		18		5		70	25	
if they be	5						7	21	
in to the							6	5	6
is good to			17		12		67	6	
is to say	10		11					7	5
it is a	5		14		7			7	5
it is called	5						27	11	
it is good			11		13		79	9	
it is not	8						5	15	
it may be	6				10		85		
it to the					12		19	5	
lay it to					5		27	5	
of man's body	12		6						12
of the body	23						9	44	13
of the liver	12						21	5	
of the same	25				12				15
of the which	16				10			19	
of the year	5		7					5	
part of the	30				5		5	6	16
that is to	9		11					30	7
the first is	5							10	8
the liver and	6						15	5	6
the parts of	5		7						7
therefore it is	7						6		5

P2 1550–1579	1	2a	2b	2c	2d	2e	3	4	5
according to the		5	6	9				7	5
and it is							8	5	48
and of the				5			14		17
by reason of		5					14	11	
if a man			8			5	10		
if it be	11		15	18		5	8	13	26
is good for			49			5	20		
is to be				7				12	42
is to say				30		6	6		63
it is a	9					6	5		13
it is called		7					5		11
it is good			50			14	27		
it is not				7		5			6
of the body		16	5				8	48	51
of the disease	8	6				5		10	
of the head			12				12		67
of the same		21	11				8		21
of the stomach							13	6	8
ought to be			8	15			8	13	11
out of the		5	20			8	23	7	5
part of the							10	18	59
that is to				31		6	6		83
that which is	5					5			12
the body and		5						15	13
the liver and							6	5	13
the time of		9						7	5
water of the			5			5	8		
when it is			7	5			6		

P3 1580–1609	1	2a	2b	2c	2d	2e	3	4	5
according to the	9	5	5		13	7		7	
and in the				6			5	7	
and such like	5					7	5	12	
are to be		5		5		5			
as it is	12	8					6		7

(*Continued*)

P3 1580–1609 (cont.)	1	2a	2b	2c	2d	2e	3	4	5
as it were	15	36					12		
as much as	6		6	8					
by reason of		5		11	7	6		9	
for the most	5					6	5	6	
if it be			16	5				10	
in our bodies				7	6		5		
in respect of		5		7			6		
in the first			15				6	5	
in the morning				6		7	16		
is to be	22	12	6	18			5		
is to say	10					7		7	
it is a			5	5		8	10		
it is good			17				13	5	
it may be	7	5		7		5			
of each a		5		13		11	9		
of the body	35	50	22	27	26	16	5	43	20
of the head		7	6				11		6
of the heart	17				5	5			
of the mind	10	16						14	
of the mother			5		36		5		
of the said		7					5		10
of the same	9	21	8	5		17	8		16
out of the	7	5	14	7		9	7		20
part of the	11	19	5				5		6
parts of the	7	5		19	11		5	5	10
that is to	25	6				8		7	
that it is	16			29		5	14	11	
that which is	20	7							9
the body and	17	12	5					13	
the cure of	9	21							10
the most part	6					8	5	6	
the nature of	12	8				5	9	5	
the stomach and			7	16			7		
the whole body	6		5		11				8
when it is	5			7			7		
with oil of			6	7			6		5

P4 1610–1639	1	2a	2b	2c	2d	2e	3	4	5
a kind of				5	5			7	
according to the				5	5			5	25
ana +q ii	7				7				10
and such like		7		5			6		11
as it were				10	7				9
because it is		6		5				7	6
by reason of	10	13		10	11			27	14
cure of the	6	8				5			7
for the most			8		6		6	5	
hot and dry		9					9	15	
if it be		6		11	24		49	7	6
in the head			10		5				5
in the morning		9			5		11	5	13
is to be		18						7	5
is very good							5	8	7
it is a			11	8	5		7		7
it may be				5	6			5	9
it shall be		8					7		7
let the patient	5						15		5
of an egg		6					9		6
of the belly					7		6		6
of the body	12	25	6		5			42	26
of the head	16	12			10				20
of the same		8		11					5
oil of roses		9					13		11
part of the	14	11						17	24
parts of the	5	12			6			5	6
reason of the				5				14	9
that it is		8	12	5			9		
that which is		7						13	13
the cure of	9	7				5			24
the juice of	7	7		5	8	6	40		
the most part			8		6		7	5	
the use of				11	11			30	
the white of		10					18		6
the whole body	8				7				8
there is a			5	8	5				
whether it be			5	6	8				
white of an		6					9		5

P5 1640–1669	1	2a	2b	2c	2d	2e	3	4	5
according to the	5		9				8	5	
and in the	7						14		17
as it were	10						5		16
by reason of	6	7	5						28
cold and dry	17						6		14
half an ounce	5				21	5	22		10
hot and dry	18		8				39		
if it be	5	9			5		21	10	25
in the body	6	7	5						9
in the head	6						10		5
in the morning	11						29	6	
in the stomach	5				7			5	
is hot and	9		5				10		
is to be			5				17		24
it is a		7					5		17
of each half					9		5		13
of each one					5		19		25
of each two					8		19		10
of the body	26	24			9		17	18	71
of the head	7						5		26
of the heart	20						6		104
of the liver	10						22		7
of the stomach	7				6			5	12
oil of roses					12		7		6
out of the		7			5		9		36
part of the	20				8		6		31
parts of the	8	11			5				24
the blood is	8	5							11
the body and	8						5		6
the head and	6				5		7		6
the liver and	6						20		7
the most part	8						5		6
the nature of	10	5	6						
the stomach and	10						6	5	7
the whole body	6				5				8
three or four	5				6		14		
when it is	6						18		5

P6 1670–1700	1	2a	2b	2c	2d	2e	3	4	5
a pint of				9			24		6
a quart of				9			5	28	
a spoonful of	6					8	11		
according to the	22	5	7	12			12	5	17
all sorts of	6						7		7
an ounce of	13			10		7	16		5
and a half	10			6		10	6		
and such like			12				6	8	
are to be	21			6			8	5	14
as much as	9					11			11
cold and moist	11		6					6	
each half a						6	7		8
each one dram	10					6			8
each two drams	6					6			5
for the most		5	5		5				10
half a dram	21					16	8		8
half a pint	6			6			11		
half an ounce	32			13		21	21		5
hot and dry	18		10				11		
if it be	22						7		13
in the beginning	7			5					9
in the body	6		11					19	
in the morning	45			18			34		10
in the same				5		5	8		
is to be	40	6		22			17	6	51
it is a			5		5		45		
it is to				5			5		8
of each half	9					11	9		10
of each one	22					15	29		14
of each three	8						7		5
of each two	9					7	13		6
of the blood		9	28	18			5		5
of the body	20	9	17	11	10		11	18	22
one dram of	10					5	8		
one ounce of	19						8		5

(*Continued*)

P6 1670–1700 (cont.)	1	2a	2b	2c	2d	2e	3	4	5
ought to be	10		12				10	5	6
ounce of the	8			9			6		
part of the	10	5	14	10	11		19		7
parts of the			5	9			6		
quantity of a	5						5		5
take of the	46					8	24		
the blood and			14	5		5			
the cause of	7		6	5					
the cure of			14	7		5	5		7
the juice of	14			8			43		
the most part		5	5		5				10
the quantity of	6	9		5			10		6
the time of	7	5	17						
the use of	17		9	27					
those that are	16					5	15		
three or four	7			10		5	42		
two or three	5	5		14		5	30	10	9
two ounces of	7					5	24		8
which is the				5			5	6	