Formulaic language in Early English Books Online: from computational linguistics to classical rhetoric

Martin Wynne
University of Oxford

Dan McIntyre
Uppsala University

Michael Burke
University of Utrecht

Introduction

As Sinclair (2004: 30) points out, "so-called 'fixed phrases' are not in fact fixed; there are very few invariable phrases in English." Consequently, any exploration of formulaic language needs to consider not only their structural forms but also their contextual functions. This insight has led to a ground-breaking study of literary style (Mahlberg 2013), based on the analysis of recurrent phrases in the works of Charles Dickens, amid an increasing interest in various forms of formulaic language. This increase of attention in clusters, or n-grams, is reflected in reflected in the widespread discussion and use of Google Ngram Viewer[1], and the inclusion of functions for n-grams in widely used corpus query applications such as AntConc (Anthony 2023) and Sketch Engine (Kilgariff et al 2014) and programming toolkits such as NLTK[2] and spaCy[3].

In this study, we analyse repeated clusters of words (n-grams) from Early English Books Online (EEBO) and subject our results to qualitative analysis using analytical frameworks from both stylistics, corpus linguistics and classical rhetoric. With regard to the latter, we investigate patterns of n-grams which correlate with rhetorical figures.

The purpose of the investigation is to apply Sinclair's hypothesis about everyday contemporary English usage to historical English as evidenced in a large collection of published books, in order to investigate to what extent it applies to the language of this corpus.

This is an initial, exploratory study, exploiting for perhaps the first time the opportunity to extract information about repeated clusters of words in such a large corpus from this period of time for the English language. The intention is to indicate avenues of future research which could make use of this newly revealed data.

The corpus: Early English Books Online

The Early English Books Online collection was created by the EEBO-TCP project. EEBO-TCP was a partnership between the commercial publishers ProQuest and with more than 150 libraries to generate accurate, fully-searchable, SGML and XML-encoded texts

[1] https://books.google.com/ngrams/
[2] https://www.nltk.org/
[3] https://spacy.io/

corresponding to books represented in the Early English Books Online collections known as Short Title Catalogues I and II (based on the Pollard & Redgrave and Wing short title catalogues respectively), as well as the Thomason Tracts and the Early English Books Tract Supplement collections. Together these trace the history of English thought from the first book printed in English in 1475 through to 1700, and include, for example, works by Erasmus, Shakespeare, King James I, Marlowe, Galileo, Caxton, Chaucer, Malory, Boyle, Newton, Locke, More, Milton, Spenser, Bacon, Donne, Hobbes, Purcell, Behn, and Defoe. The books in these collections include works of literature, philosophy, politics, religion, geography, history, politics, mathematics, music, the practical arts, natural science, and other areas of human endeavour as evidenced in printed works of the period. The assembled collection of more than 125,000 volumes is central to understanding the development of Western culture, and the Anglo-American world in particular. The STC collections have perhaps been most widely used by scholars of English, theology, linguistics, and history, but these resources also include core texts in art, women's studies, history of science and medicine, law, and music. The selection and encoding of texts is reviewed critically by Gadd (2009).

The repeated clusters of words generated and analysed in this paper were extracted from the selection of the entire EEBO collection which is now in the public domain, as curated by the Oxford Text Archive. The corpus consists of 60,328 printed works, with approximately 1.5 billion words. Our study therefore constitutes what is likely to be the largest extraction of n-grams from a historical corpus of English to date. The n-gram frequency lists will be made freely available online via the Oxford Text Archive[4].

This sub-section of the total EEBO database represents the outputs of Phase 1 and Phase 2 of the EEBO-TCP project. From 2000-2009, Phase I successfully converted 25,368 selected texts from the Early English Books Online corpus. Phase I texts were initially available only to institutions that contributed to their creation, but were then released to the general public in 2015, and are therefore currently available for access, distribution, use, or reuse by anyone.

Phase 2 produced a further 34,963 texts, made freely available to the public in 2020 according to a Creative Commons (CC0) licence by project partners in University of Michigan[5] and the University of Oxford[6].

Selection of works to transcribe for EEBO-TCP was based on a number of factors, including inclusion of named authors mentioned in the New Cambridge Bibliography of English Literature to ensure selection of canonical, or at least attributed, works. The corpus includes

---

[4] http://hdl.handle.net/20.500.14106/2547
[5] https://quod.lib.umich.edu/e/eebogroup/
[6] http://hdl.handle.net/20.500.14106/5

works related to a wide variety of fields, not just literary studies. In a later stage, works exemplifying particular themes (e.g. food, drugs, piracy, witchcraft), or fitting a particular format (broadsides, pamphlets, etc.), were selected.

The specific version of the collection used in this project was the entirety of the texts available in the Oxford Text Archive collections from the University of Oxford, numbering 60,328 works and more than 1.5 billion words (tokens, excluding punctuation) in total. The version of the text used was that prepared for the SAMUELS Project[7], in one-word-per-line format with normalized orthography, annotation for part-of-speech, lemma and semantic fields (see fig.1 for an example).

```
<document LANG='eng'>   XMLCODE NULL            0                               Z99     NULL    04:10
<FRONT> XMLCODE NULL            0                               Z99     NULL    04:10
<P>     XMLCODE NULL            0                               Z99     NULL    04:10
A       a       AT1             0                       Grammatical     Z5      04:03   ZC
Letter  letter  NN1     Q3      0       03:09:07:06-02 02:01:14:09:05:06-04     BG:07:f AR:52:b:01      letter  Q1:2    03:09:07:04-01  BG:07:d
written write   VVN     I2:1    0       03:09:07-04 02:07:05-01 BG:07 AX:21     Write   Q1:2    03:09:07        BG:07
out     out     II21    A9-     1:2:1   01:12:06:01-06:01 01:12:06:01-06       AL:06:a M       Z5      01:14:05:11     AN:05:k
of      of      II22    A9-     1:2:2   01:12:06:01-06:01 01:12:06:01-06       AL:06:a M       Z5      01:14:05:11     AN:05:k
England england NP1             0                               Z2      04:01:02        ZA02
to      to      II              0                       Grammatical     Z5      04:03   ZC
an      an      AT1             0                       Grammatical     Z5      04:03   ZC
English english JJ      Z2/Q3 Z2/S2mfnc 0       04:01:02 04:01:02       ZA02 ZA02               Z2      04:01:02        ZA02
Gentleman       gentleman       NN1             0       03:01:06:01:01:01-02 02:02:09:05:05:01-03:02    AY:06:a:01 AS:13:c      gentleman       S2:2m   03:01:06:01:03:01-08    AY:06:a:03:a
remaining       remain  VVG     M8 N5:2+        0       02:01:14:01 01:12:04:03-03      AR:50 AL:04:c           T2++    01:16:07:04:01:01       AP:07:d:01:a
at      at      II              0                       Grammatical     Z5      04:03   ZC
Padua   padua   NP1             0                               Z99     NULL    04:10
,       PUNC    YCOM            0                               PUNC    NULL    NULL
containing      contain VVG     A1:7+   0       01:12:05:01:01-05 01:13:02-01   AL:05:a AM:02   endure/remain/persist/continue  A1:8+   01:12:05:01:01  AL:05:a
a       a       AT1             0                       Grammatical     Z5      04:03   ZC
true    true    JJ      A5:4+   0       01:12:06-02:01 03:13:03:04:07:04:01-03  AL:06 BK:06:g:03        unchangeable    A5:2+   01:13:11:05-02  AM:11:e
Report  report  NN1     X3:2    0       01:09:09:02:03:03-07:02 AI:15:c:03              Q2:2    03:09:05:04-02  BG:05:d
of      of      IO              0                       Grammatical     Z5      04:03   ZC
a       a       AT1             0                       Grammatical     Z5      04:03   ZC
strange strange JJ              0       02:01:12:07:05 01:16:03:02:02:01:01-04:02       AR:37:d AP:03:b:02:a    strange A6:2-   01:10:07:05:04:07-05    AJ:07:e:03:d
Conspiracy      conspiracy      NN1     G2:2-   0       02:05:04:01:01-02 03:05:09:02:01-20     AV:03:a:01 BC:06:b      conspiracy      G2:1-   02:05:04:01:01-03:03    AV:03:a:01      Conspiracie
,       PUNC    YCOM            0                               PUNC    NULL    NULL
contrived       contrive        VVN     X9:2+   0       01:10:10:02-09:01 01:11:02-04   AJ:10:b AK:02   fashion/shape/form      X7+     01:11:02-06     AK:02   contriued
between between II              0                       Grammatical     Z5      04:03   ZC      betweene
Edward  edward  NP1             0                               Z1m     04:01:01        ZA01
Squire  squire  NN1     S2:2m   0       03:01:06:01:01:01-07 01:15:21:04:01:02:02-03    AY:06:a:01 AO:22:c:01   Squire  S7:1+/S2:2m     03:01:06:01:02:09:01    AY:06:a:02:f
,       PUNC    YCOM            0                               PUNC    NULL    NULL
lately  lately  RR              0       01:15:20:06 01:14:04:02-04      AO:21:f AN:04:b Unwillingly     T3---   02:05:03:02     AV:02:a:01
execed  exece   VVD             0                               Z99     NULL    04:10
for     for     IF              0                       Grammatical     Z5      04:03   ZC
the     the     AT              0                       Grammatical     Z5      04:03   ZC
same    same    DA              0       01:16:01:04-02:01 01:16:01:04-02:02     AP:01:d AP:01:d Same    A6:1+++ 01:16:01:04     AP:01:d
treason treason NN1             0       02:03:05:15:01 03:06:01:03:02:03-08     AT:16 BD:01:c:02:c      instance        G2:1-   02:03:05:15:01-03       AT:16
```

Figure 1: Fragment of a SAMUELS version of an EEBO text[8]

Extracting N-grams from EEBO

The SAMUELS versions of the EEBO texts were processed with a python script, using the NLTK *ngrams* function and linux shell commands, to produce ranked, ordered lists of the most frequently occurring sequences of words. The tab-separated format of the SAMUELS files (see figure 1) allowed the selection of relevant columns - e.g. column 1 for the orthographic word. It should be noted that in the SAMUELS versions of the texts historical spelling variants have been changed to a normalized version by the VARD variant detector application[9] (see, for example, *Conspiracy/Conspiracie* in Figure 1). Further columns in the SAMUELS files contain annotations relating to semantic fields and links to historical thesaurus entries, and were not used in this study.

The following procedures were therefore carried out in the following order:

[7] https://www.gla.ac.uk/schools/critical/research/fundedresearchprojects/samuels/
[8] http://hdl.handle.net/20.500.14106/A00021
[9] https://ucrel.lancs.ac.uk/vard/about/

- extract the relevant data column (e.g. column 1 for the word token; column 2 for lemma; column 3 for POS-tag, etc.);
- filter out unwanted lines (e.g. XML tags representing metadata or structural units; punctuation for step 02);
- convert lines where necessary (e.g. convert upper-case to lower-case; convert all numbers to 'NUM');
- run the NLTK ngrams program to extract all n-grams (one iteration each for 4-, 5- and 6-grams);
- sort, count, and sort again in descending frequency order (with linux shell scripts) to extract the top 1000 clusters at each stage.

This procedure was carried out 19 times with different parameters in order to create ranked frequency lists of n-grams for differing values of word, calculated in the following versions. The justification for these steps is discussed below the section 'Preliminary Results'.

- Step 1: Just the words (all tokens, including punctuation, case sensitive)
- Step 2: Just the words (minus punctuation, case sensitive)
- Step 2i: Just the words (minus punctuation, case sensitive)
- Step 3: Lemmas
- Step 4: Lempos (lemma and part-of-speech pairs)
- Step 5: Part-of-speech (POS) tags
- Step 6: Phrase frames, lemmas with wildcards in positions 2,3,4

For steps 1-5, 4-, 5- and 6-grams were generated, producing 18 lists in total. For step 6, which requires significantly greater computer processing power and time, only 5-grams were generated.

In order to allow inspection, verification and reproducibility of our research, and to allow other researchers to make use of the data, the full lists are made publicly available with Creative Commons licence[10] via the Oxford Text Archive collections from the UK CLARIN repository[11].

Preliminary Results

Analysis of the most frequent n-grams reveals a preponderance of formulaic expressions from religious texts, often bible quotations or popular prayers and other ritual expressions. Analysis of the rhetorical and stylistic functions of these clusters allows insight not only into the nature of formulaic expression in Early Modern English but also the nature of

[10] https://creativecommons.org/licenses/by-nc-sa/4.0/
[11] http://hdl.handle.net/20.500.14106/2570 (to be published January 2024)

interpersonal communication in this period. In this respect, our study demonstrates the capacity of large corpora of written material to offer insights into the nature of communication across both writing and speech.

EEBO 4-grams, just the (orthographically normalized) words, case-sensitive (step01)

freq	*ngram*
233273	, &amp ; c :
52649	that is to say ,
40787	, that is to say
24367	, by reason of the
22861	the word of God ,
22761	of God , and the
20744	, that is , the
20605	of the Church of England
20123	&amp ; c : And
19851	the Son of God ,
18813	of the Church , and
18780	, as it were ,
18232	, as well as the
17936	&amp ; c : The
17924	the Church of England ,
17834	on the other side ,
17456	of the Lord , and
17408	the Church of Rome ,
16590	the Word of God ,
16500	: In the mean time

Figure 2: Step 1 4-grams

As can perhaps be seen from Figure 2, inclusion of punctuation has a strong effect on the results, with all of the twenty most frequent patterns including at least one punctuation character. In order to focus more closely on lexical choices in texts, the analysis was run again (step 2) with punctuation filtered out before the calculation of n-grams. The results for 4-grams (Figure 3) and 5-grams (Figure 4) are shown.

While the exclusion of punctuation helped considerably with the identification of repeated clusters of words, there are still some issues. The most obvious was the inclusion of similar clusters as separate entries where capitalization differs, for example "the word of God" and "the Word of God" (Figure 3). A variant of step 2 in which all upper-case characters were converted to lower-case, was therefore carried out, named step2i (i for case insensitive).

EEBO 4-grams (orthographic words, case insensitive)

freq	*ngram*
103063	the word of god
78394	that is to say
65964	the rest of the
65815	in the time of
62735	the son of god
62071	in the midst of
60945	the end of the
60171	in the mean time
58377	of the church of
54988	for the most part
54800	the spirit of god
54767	the name of the
54011	the church of rome
51984	by reason of the
50493	in the name of
48316	at the same time
48054	it can not be
47918	on the other side
47320	the church of england
46960	is not to be

Figure 3: Step 2i 4-grams with all words converted to lower case

As a further attempt to see beyond variations in inflected forms, a further step was carried out using lemmas rather than orthographic words (step 3). The lemmas, as assigned by researchers in the SAMUELS project, are available in the source files. In order to account for an examine the use of numbers in texts, a version was run with all numbers converted to NUM, revealing that the most frequent clusters were sequences of numbers, and a version without, in order to allow the closer examination of lexical choices. Results are shown in figures 5 and 6.

EEBO 4-grams, just the (orthographically normalized) words, punctuation pre-filtered out:

freq	*ngram*
68543	that is to say
62411	the rest of the
59958	in the time of
59165	in the midst of
56335	the word of God
54627	of the Church of
54118	for the most part
51833	the end of the
50574	by reason of the
47904	the Church of Rome
46486	is not to be
46449	the Son of God
44245	the Church of England
42510	at the same time
41491	it can not be
39242	the time of the
38717	on the other side
38097	the name of the
38064	it is to be
37757	the Word of God

Figure 4: Step 2 4-grams

EEBO 5-grams, just the (orthographically normalized) words, punctuation pre-filtered out

freq	*ngram*
20623	of the Church of England
16259	in the midst of the
15800	of our Lord Jesus Christ
14770	in the sight of God
14729	to the end of the
14597	in the name of the
13590	in the time of the
13302	of the Church of Rome
11644	the end of the world
11315	and the rest of the
11064	the year of our Lord
10960	the name of the Lord
10927	of the Son of God
10400	it is not to be
9696	in the beginning of the
9373	that is to say the
9133	la la la la la
8799	the one and the other
8733	all the rest of the
8433	at the right hand of

Figure 5: Step 2 5-grams (top 20)

EEBO 5-gram lemmas, no punctuation, all numbers converted to 'NUM'

freq	*ngram*
1254278	NUM NUM NUM NUM NUM
27991	NUM NUM and NUM NUM
24609	NUM NUM & NUM NUM
23119	in the name of the
21736	of the church of england
19301	the name of the lord
18936	the end of the world
17501	NUM cor NUM NUM NUM
17302	in the time of the
17044	in the midst of the
16968	to the end of the
16742	NUM NUM & NUM NUM
16648	NUM NUM NUM NUM and
16592	the word of god and
16478	of our lord jesus christ
16263	it be not to be
16165	of the son of god
15960	in the sight of god
15830	be the son of god
15359	the word of the lord

Figure 6: Step03 5-grams (all numbers converted to 'NUM')

When the most frequent 5-grams were calculated, it was discovered that repeated sequences of numbers tended to dominate the top of the list. It was decided to make a version of this list without numbers (by filtering out all lines containing at least one number),

since this was mainly an artefact of the conversion of all numbers to the NUM tag.

EEBO 5-gram lemmas, no punctuation, all lines containing numbers removed

freq	*ngram*
23119	in the name of the
21736	of the church of england
19301	the name of the lord
18936	the end of the world
17302	in the time of the
17044	in the midst of the
16968	to the end of the
16592	the word of god and
16478	of our lord jesus christ
16263	it be not to be
16165	of the son of god
15960	in the sight of god
15830	be the son of god
15359	the word of the lord
14677	the year of our lord
14354	of the church of rome
12901	the body and blood of
12874	of the word of god
12137	as it be in the
11831	and the rest of the

Figure 7: Step03 5-grams (without numbers)

It was decided that the Step03 5-grams (case insensitive, without punctuation or numbers), as seen in figure 6, provided the best starting point for more detailed, qualitative investigation of formulaic language. The analysis in the sections below relate to these n-grams.

For completeness, the results for the most frequent clusters in steps 4-6 are also shown below, but are not discussed further in this paper.

EEBO 5-grams for 'lempos' (lemma and part-of-speech pair)

freq	*ngram*
22814	in_II the_AT name_NN1 of_IO the_AT
21477	of_IO the_AT church_NN1 of_IO england_NP1
17204	the_AT end_NN1 of_IO the_AT world_NN1
17044	in_II the_AT midst_NN1 of_IO the_AT
15907	of_IO our_APPGE lord_NNB jesus_NP1 christ_NP1
15864	the_AT word_NN1 of_IO god_NP1 and_CC
15392	to_II the_AT end_NN1 of_IO the_AT
15165	in_II the_AT sight_NN1 of_IO god_NP1
14913	in_II the_AT time_NNT1 of_IO the_AT
14589	the_AT name_NN1 of_IO the_AT lord_NN1
14200	of_IO the_AT church_NN1 of_IO rome_NP1
14195	it_PPH1 be_VBZ not_XX to_TO be_VBI
13306	of_IO the_AT son_NN1 of_IO god_NP1
12943	the_AT year_NNT1 of_IO our_APPGE lord_NN1
12890	the_AT body_NN1 and_CC blood_NN1 of_IO
11830	and_CC the_AT rest_NN1 of_IO the_AT
11263	of_IO the_AT word_NN1 of_IO god_NP1
11238	the_AT word_NN1 of_IO the_AT lord_NN1
11212	in_II the_AT beginning_NN1 of_IO the_AT
10512	body_NN1 and_CC blood_NN1 of_IO christ_NP1

Figure 8: Step04 5-grams

EEBO 5-posgrams (part-of-speech tags only, step05)

freq	*ngram*
2079655	AT NN1 IO AT NN1
1291507	II AT NN1 IO AT
1139407	II AT NN1 IO NP1
1118100	MC MC MC MC MC
1004375	NN1 II AT NN1 IO
925307	AT NN1 IO APPGE NN1
806335	NN1 NN1 NN1 NN1 NN1
780550	AT NN1 CC NN1 IO
640550	VVN II AT NN1 IO
628851	NN1 IO AT NN1 IO
601239	II AT NN1 IO NN1
597486	NN1 IO AT JJ NN1
568283	AT NN1 IO AT JJ
559385	II AT NN1 IO APPGE
535544	II AT JJ NN1 IO
532875	AT NN2 IO AT NN1
525327	AT NN1 IO AT NN2
518563	NN1 IO AT NN1 CC
510266	NP1 NP1 NP1 NP1 NP1
493685	AT NN1 IO NP1 CC

Figure 9: Step05 5-posgrams

EEBO 5-phrasegrams, excluding all lines which included numbers (step06b)

freq	*ngram*
431741	in the # of the
398333	of the # of the
363001	to the # of the
231677	be the # of the
214887	by the # of the
205596	the # of the lord
195590	for the # of the
188647	and the # of the
167684	the # of god and
149426	the # of the church
141866	the # of the world
138775	of the # of god
135234	of the # and the
129711	from the # of the
113775	in the # of god
112567	that the # of the
104029	all the # of the
101868	with the # of the
98929	it be the # of
98772	at the # of the
97670	in the # of his
97254	out of the # of
91810	be the # of god
91457	be in the # of
90776	to the # of god
88738	the # of god be
88369	the # part of the
78497	of the # in the
77062	the # of the gospel
75213	the # of the law

Figure 10: Step06 5-phrasegrams (with numbers filtered out)

A perspective from stylistics

Intuitively, it would seem that the majority of the 4164 5-grams are related in some way to the concept of religion. To get a sense of the proportion of n-grams which expressed religiosity, we examined the first 1000 and found 289 (i.e. 28.9% of our sample) to contain explicit references to religion (e.g. to God, Jesus Christ, the Church of England, etc.). In addition to these, a large proportion of the 1000 n-grams that we sampled are also likely to be references to religion and/or the Bible, though we did not include these in our initial count since these are only plausible/potential references. Examples of these include ENTER INTO THE KINGDOM OF, OF THE LAND OF EGYPT and BODY AND BLOOD OF OUR.

The most frequent 5-gram is the prepositional phrase IN THE NAME OF THE (23,119 hits). Since the size of the database currently makes qualitative analysis prohibitive, we examined this phrase in EEBO (Version 2) using CQPWeb. Our search returned 7892 hits

across 2355 different texts. The phrase has a relative frequency of 12.642 instances per million words. Figure 1 shows the top 50 collocates of IN THE NAME OF, with a log ratio of at least 5 (meaning the collocate is at least 32 times more frequently found in proximity to the node – i.e. the 5-gram – than it is elsewhere). Of these collocates, at least 30 may be ascribed to a semantic field of religion. In order to investigate the functions of the 5-gram in more detail, we examined the context of the first 100 of these hits and found that 61 of the referents of the phrase were religious while 39 were secular. The most frequent religious referent was FATHER (28 hits, incorporating instances of FATHER, SON AND HOLY GHOST) and LORD (i.e. IN THE NAME OF THE LORD), with 24 hits. The other religious referents were BLESSED TRINITY (3), HOLY TRINITY (1), TRINITY (1), SAVIOUR OF THE WORLD (1), CHURCH (1), GOD OF ISRAEL (1) and APOSTLES (1). Of the secular referents, the majority are either royalty (FRENCHE KYNG, KYNGE OF FRAUNCE, KYNGE OF ENGLANDE), aristocracy (e.g. DUCHESSE, LADY WYFE TO THE LORD CHARLES OF BLOYES, DUKE OF ANIOU) or political entities (e.g. HOLE COUNTRE OF FLAUNDERS). That is to say, all of the referents of the 5-gram fulfil high-status societal roles. In stylistic terms, this points towards the function of the 5-gram, which in many cases is to serve as explicit reference to the felicity conditions of performative speech acts (e.g. I BAPTYSE THE IN THE NAME OF THE FATHER, THE SONE AND OF THE HOLY GHOOST, AND THE FORSAYD MONKES CONIURED HYM IN THE NAME OF THE TRYNYTE TO GOO THENS and IN THE NAME OF THE FATHER & SONNE AND HOLY GHOSTE I HENRY OF LANCASTRE). Austin (1962: 14-15) notes that for a performative to be recognized as such, ‘the circumstances and persons must be appropriate’ and ‘[t]he procedure must be executed (i) correctly, (ii) completely’. In effect, the 5-gram provides evidence that this is the case, by offering explicit reference to the authority on whom the speaker’s words rest.

By way of diachronic contrast, there are 198 instances of IN THE NAME OF THE in the original BNC. Of these, a smaller proportion (21%) of the referents are religious than in EEBO, with the majority being secular. While our qualitative analysis so far has been of a relatively small amount of data, these findings suggest interesting hypotheses and research questions for further investigation. For instance, does the decline in religious referents for the 5-gram IN THE NAME OF THE reflect an increasingly secular society, or is it more a case that religiosity is now expressed outside the structures of formal religious texts? Does IN THE NAME OF THE have a function beyond performative speech acts? And are there any contexts in which IN THE NAME OF THE collocates with referents that are not explicitly religious, royal, aristocratic or political?

No.	Word	Total no. in whole corpus	Expected collocate frequency	Observed collocate frequency	In no. of texts	Log Ratio (filtered)
1	Editors	80	0.006	7	3	10.304
2	baptising	622	0.047	48	42	10.106
3	baptise	1,412	0.107	84	63	9.704
4	baptyse	115	0.009	6	5	9.503
5	baptizing	4,461	0.338	221	168	9.424
6	Hosanna	768	0.058	34	25	9.254
7	cummeth	307	0.023	11	1	8.936
8	trynyte	257	0.02	8	6	8.726
9	adjured	174	0.013	5	5	8.607
10	baptize	6,246	0.474	154	108	8.38
11	baptysed	546	0.041	11	9	8.082
12	Adjure	415	0.032	6	6	7.595
13	Baptiz'd	886	0.067	12	6	7.5
14	Lord]	559	0.042	7	5	7.385
15	baptised	4,978	0.378	62	47	7.377
16	Trinitie	3,845	0.292	44	34	7.254
17	Baptized	25,195	1.911	274	145	7.179
18	Praeses	744	0.056	8	2	7.163
19	Trinity	13,882	1.053	126	69	6.916
20	Hoasts	647	0.049	5	4	6.682
21	coniure	717	0.054	5	5	6.533
22	uncreated	734	0.056	5	3	6.499
23	cometh	20,180	1.531	132	67	6.44
24	speakest	1,531	0.116	10	9	6.437
25	Zeph	1,080	0.082	7	7	6.426
26	justified	21,662	1.643	135	86	6.369
27	Hosts	5,261	0.399	32	30	6.334
28	Godfathers	937	0.071	5	5	6.144
29	Father	269,913	20.473	1,321	617	6.019
30	oil	3,115	0.236	15	11	5.995
31	begotten	14,220	1.079	67	61	5.964
32	✗	2,651	0.201	12	8	5.906
33	salutes	1,113	0.084	5	3	5.895
34	Goliah	1,404	0.107	6	5	5.822
35	Commonalty	1,416	0.107	6	4	5.81
36	Jesus	79,979	6.067	335	199	5.793
37	Lord	624,276	47.352	2,431	1023	5.688
38	prophesied	3,816	0.289	14	12	5.601
39	blesse	13,812	1.048	49	39	5.553

40	commeth	30,327	2.3	107	70	5.545
41	dip	1,419	0.108	5	5	5.543
42	Infanta	1,430	0.109	5	3	5.532
43	oile	4,040	0.306	14	10	5.519
44	washing	7,181	0.545	24	17	5.466
45	pronounceth	2,097	0.159	7	7	5.464
46	commons	28,366	2.152	92	51	5.423
47	lorde	75,570	5.732	245	80	5.422
48	co~meth	2,822	0.214	9	8	5.398
49	Iesus	83,826	6.358	263	171	5.375
50	[I	2,573	0.195	8	8	5.362

Figure 11 Collocates of IN THE NAME OF, using a 5-word window either side of the node

Conclusion

As is indicated in the introduction, this is an exploratory study, focussed on extracting new information about repeated clusters of words in a large historical corpus of English published works. The preliminary analysis offered in this paper reveals some of the functionality of the n-grams, notably that they often fulfil the requirements associated with the performance of specific speech acts. This analysis demonstrates how large corpora can be used to gain insights not only into the structural aspects of earlier Englishes but the pragmatic aspects too. In the absence of contemporary spoken language samples from the period, this method also offers an important means of accessing information about interpersonal elements of communication. We suggest that directions for future research could include a more systematic analysis of distributions of clusters over time and across text types, further analysis of the stylistic and pragmatic functions of n-grams, the further investigation of rhetorical figures as evidenced in the n-grams, comparison of EEBO n-grams with other text types and time periods, and the investigation of the sources of n-grams, where they are intertextual references.

Bibliography

Anthony, Laurence, *AntConc (Version 4.2.4)*, Tokyo, Japan: Waseda University, 2023, https://www.laurenceanthony.net/software.

Austin, J. L., *How to Do Things With Words*, Oxford: Clarendon Press, 1962.

Gadd, Ian, "The Use and Misuse of Early English Books Online", *Literature Compass,* 6:3, May 2009: 680-692.

Kilgarriff, Adam, Baisa, Vít, Bušta, Jan, Jakubíček, Miloš, Kovář, Vojtěch, Michelfeit, Jan, Rychlý, Pavel, Suchomel, Vít, "The Sketch Engine: ten years on", *Lexicography*, 1: 7-36, 2014.

Mahlberg, Michaela, *Corpus Stylistics and Dickens's Fiction*, London: Routledge, 2013.

Piao, Scott, Fraser Dallachy, Alistair Baron, Jane Demmen, Steve Wattam, Philip Durkin, James McCracken, Paul Rayson, Marc Alexander, "A Time-Sensitive Historical Thesaurus-Based Semantic Tagger for Deep Semantic Annotation," *Computer Speech & Language* (46), 2017.

Rayson, Paul, Archer, Dawn, Baron, Alistair and Smith, Nick, "Tagging historical corpora - the problem of spelling variation," *Proceedings of Digital Historical Corpora*, Dagstuhl-Seminar 06491, International Conference and Research Center for Computer Science, Schloss Dagstuhl, Wadern, Germany, 3rd-8th December 2006. ISSN 1862-4405.

Sinclair, John, *Trust the Text,* London: Routledge, 2004.