Detecting Formulaic Language Use in Historical Administrative Corpora

Marijn Koolen[1,2], Rik Hoekstra[1,2]

[1]KNAW Huygens Institute, Amsterdam, the Netherlands
[2]DHLab, KNAW Humanities Cluster, Amsterdam, the Netherlands

Abstract
Historical administrative corpora are filled with jargon and formulaic expressions that were used consistently across many documents. Governmental decisions, notarial deeds and official charters often contain fixed expressions to ensure that the same legal aspects in different documents had the same interpretation. Such formulaic expressions can be used to identify specific elements of a document. For instance, a deed has different formulas to indicate whether it concerns the sale of property or the transferal of rights. In this paper we explore formulas as a methodological devise to structure the text of an administrative corpus and make the information contained in it better accessible. We use a data-driven method to detect potential formulaic expressions in historical corpora, that can deal with spelling variation and change and recognition errors introduced in the digitisation process. We apply this exploratory technique on a corpus of almost 300,000 eighteenth-century resolutions of the States General of the Dutch Republic and find many formulaic expressions that capture relationships between the political actors involved and the decisions that were made. A first analysis suggests that many formulas can be used to add metadata to individual resolutions on various elements of the proposals and decisions that are part of each resolution.

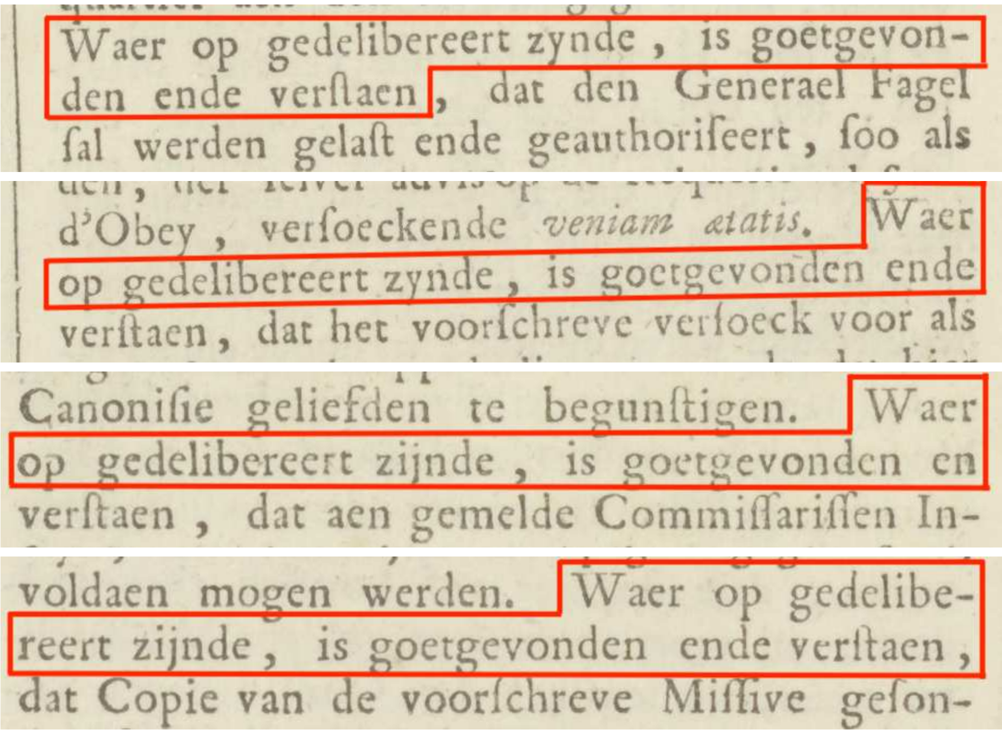
Keywords
formulaic expressions, text reuse, document structure, information extraction, text analysis

1. Introduction

The Resolutions of the States General of the Dutch Republic (1576-1796) is a digitised archive containing an estimated 1 million decisions made by the States General (SG) during their daily meetings. It is filled with administrative jargon and formulaic expressions that were used consistently, tens of thousands of times across a 220 year period, in resolutions with a very fixed structure. These formulaic expressions were used to signal specific elements in the text, so that anyone relying on the resolutions for their day-to-day work could easily find back requests, decisions and agreements by looking for these fixed phrasings, which also made sure that similar decisions and agreements had similar interpretations.

In this paper we explore formulas as a methodological device to structure the text of the archive and make the information contained in it better accessible. The secretaries of the meetings used a fixed structure and fixed expressions to signal the opening of a new resolution,

CHR 2022: Computational Humanities Research Conference, December 12 – 14, 2022, Antwerp, Belgium
marijn.koolen@gmail.com (M. Koolen); rik.hoekstra@di.huc.knaw.nl (R. Hoekstra)
https://marijnkoolen.com/ (M. Koolen)
0000-0002-0301-2029 (M. Koolen); 0000-0002-6951-8014 (R. Hoekstra)



Waer op gedelibereert zynde , is goetgevonden
ende verſtaen , dat den Generael Fagel
ſal werden gelaſt ende geauthoriſeert , ſoo als

d'Obey , verſoeckende *veniam ætatis.* Waer
op gedelibereert zynde , is goetgevonden ende
verſtaen , dat het voorſchreve verſoeck voor als

Canoniſie geliefden te begunſtigen. Waer
op gedelibereert zijnde , is goetgevonden en
verſtaen , dat aen gemelde Commiſſariſſen In-

voldaen mogen werden. Waer op gedelibereert
zijnde , is goetgevonden ende verſtaen ,
dat Copie van de voorſchreve Miſſive geſon-

Figure 1: The formulaic expression 'Waer op gedelibereert zijnde, is goetgevonden ende verstaen' used in four resolutions taken from a single meeting on the 22nd of November 1709.

that always started with a proposal submitted to the SG. Each resolution ends with a decision paragraph, which also starts with a formulaic expression, followed by the details of what was agreed upon and what should happen next. For instance, to signal the SG reached an agreement on what should be done in response to a proposition, they used the formula 'Waer op gedelibereert zijnde, is goetgevonden ende verstaen, …' (EN: *On which has been agreed and understood* …. A number of examples of this phrase are shown in Figure 1). This phrase recurs tens of thousands of times in the resolutions and signals that an agreement and decision were reached that are detailed in the following paragraph.

The formulas thus not only help us structure the material, but also to add metadata to the individual resolutions and make them better accessible for analysis.

Our experience in working with other historical collections prompted a set of questions: Are such formulaic expressions used in other administrative corpora? And what other domains and document genres contain formulaic expressions? From an information access perspective, is the textual repetition of an administrative corpus like the Resolutions different from textual repetition in corpora in other domains?

To get a better idea of the relevance of formulas, we first compare repetitive text characteristics in corpora of different domains, to establish whether administrative texts like the resolutions are of a different quality from other types of text. The second topic of this paper is the methodology we use for algorithmically identifying potential formulas in the resolutions, and a discussion of their use. In this paper we confine ourselves to these points, but we realise that formulas have relevance for a number of wider humanities research questions. We discuss these at the end in Section 6.

2. Related Work

Our work touches on three strands of research: 1) formulaic language use, 2) text reuse detection and 3) dealing with variation-rich text.

2.1. Formulaic Language

The use of formulaic expressions is mostly studied in the fields of linguistics [47, 26] and language learning [14, 7, 42, 13]. Formulaic language is typically defined as fixed word combinations, with often non-literal meanings, that are used to improve fluency and reduce misunderstanding [47, 41, 46]. Poß and Wouden studied formulaic expressions consisting of anything more than one word as *Extended Lexical Units*, stored as single a entry in a speaker's mental lexicon.

We found two studies that investigated the use of formulas in text corpora. Karsdorp identified and classified formulaic opening and closing expressions of Dutch folk tales. Repetitive patterns in the first and last five words of folk tales are detected and are found to be predictive of the genre of a folk tale. That is, the opening formula often signals that something is a joke, saga or fairy tale. In the resolutions, the opening formulas are similarly indicative of what kind of proposal (petition, report, declaration, etc.) is discussed in the resolution. [37] manually annotated the use of formulaic expressions in seventeenth and eighteenth century Dutch letters and found that more experienced writers tend to use more fixed expressions. They suggest that this indicates that formulas are partly used to reduce cognitive effort.

2.2. Text Reuse Detection

Text reuse detection has been studied extensively in the context of plagiarism detection. The annual PAN competitions, starting in 2009, have been a main driver for developing algorithms for plagiarism detection and text reuse detection [29, 27, 28, 43, 24].

Most research on text reuse detection focuses on modern texts, often digital born and using modern language, which has rules for spelling and syntax. Detecting text reuse becomes more complicated for long serial archives covering historic documents from an extensive historical period in which the language used had no consistent spelling, spelling changed over time, and the digitisation of those documents introduces text recognition errors [45, 44].

In addition to plagiarism detection, textual repetition has been studied extensively in the context of text alignment, collation and comparison [10, 40, 15], and text reuse [45, 38]. But there is remarkably little previous work focusing on the identification of formulaic expressions. We found several digital humanities studies regarding structure in text in general [1, 2, 31, 36, 45, 38, 39] and some more specific studies for technical text [8] and legal arguments [32, 11, 12]. In all these cases, the object of study is text *repetition* and the use of isolated specific terminology (noun phrases) rather than the use of *formulaic phrases.* To the best of our knowledge, outside of linguistics, humanities scholars have not written much about formulaic language use. Probably, only serial use of textual features make formulas useful for study. Scholars who have to read through them tend to see them as repetitive textual features without relevant information content.

2.3. Issues with variation-rich text

One of the big challenges of text analysis on corpora of historical texts is that they are rich in spelling variation. Many historical languages had no standard spelling and changed in spelling over time. Moreover, many texts extracted from digitised documents contain text recognition errors. These issues together lead to possibly many different spellings of the same word or phrase.

This challenge can to some extent be addressed by normalising the spelling. This maps spelling variants of words to a standard or ‘normal’ spelling of the word. VARD2 [5, 16] is a lexicon-based technique that was originally developed for historical English but ported to different historical languages. TICCL [34, 35] was developed originally to automatically normalise very large collections of 19th and 20th century Dutch. A different approach is to use fuzzy string matching and searching starting from a list of known phrases [22]. There are recent techniques based on deep neural networks, like PIE [23], that can be trained to lemmatise variation-rich languages, resulting in ‘normalised’ lemmas. However, this also reduces morphological variation that can be meaningful in distinguishing between expressions. Another drawback is that this requires a large amount of training material of linguistically annotated text.

Since we are using the same corpus as [22], we took inspiration from their fuzzy searching approach, but since it requires knowing the formulas in advance and the technique becomes very slow when a large number of formulas is used for searching, we decided to use a simplified approach of detecting common word n-grams and using character n-gram indexing to find orthographically similar spellings.

3. Formulaic expressions and their use

We narrow down our object of research with a more precise but still pre-theoretical definition of formulaic expressions and their context. A literature search has not given us any definition of formulaic expressions beyond the notions of *lexical bundles* and *idiomatic expressions* in common language use [7, 26]. Lexical bundles are often noun phrases and examples of domain specific terminology. We take a information theoretical perspective, and need a definition that is applicable to different corpora and that helps to identify formulaic expressions from large amounts of text. Given the nature of the corpus of resolutions and the corpus-specificity of its formulaic expressions, we need a definition that takes into account that formulas tend to be longer phrases, though not necessarily complete clausal units, that can incorporate and give context to variable elements like names of persons, organisations and locations or dates.

3.1. Characteristics of Formulaic Expressions

We define *formulaic expression* as a multi-word phrase (an *extended lexical unit*) that *is reused often across documents in a collection*, with *minimal word variation*, but with *potentially high variation in spelling*.[1] They may contain variable elements, spans consisting of e.g. entity names or dates. In the resolutions, a phrase might express that a certain type of proposal was submitted by a person, whose name is a variable element in the formula. As far as we can tell, this definition captures the formulas found by Rutten and Wal as well. What constitutes a formulaic expression might differ across domains, genres or corpora. We will discuss this further at the end of this paper in Section 6, but note here that this definition does not yet give us proper criteria for deciding what is a formula and what is not. Therefore, our research design is exploratory, rather than descriptive or explanatory [4, pp.91-92]. It serves to give us a better understanding of the phenomenon of formulaic expressions and how we can develop methods to study them, rather than to offer a precise description of how they are used or an explanation of how they emerge or evolve.

The study of formulaic language use has identified a number of reasons for speakers and authors to use formulaic expressions [42]. Within the domain of legal and administrative texts, the most relevant are precision of communicated information (e.g. "Cleared for takeoff" signals permission to enter a runway and commence takeoff), and signalling the structure of discourse (e.g. "on the other hand" signals an opposition) [13, p.46].

The formulaic expressions that are used over and over in many historical legal and administrative documents, serve as precise referents to the information that the document is to communicate. They prevent differences in interpretation, but also structure the information of the document. For instance, in a corpus of proclamations, a standard opening phrase signals where each proclamation starts. Notarial deeds often have template text to indicate the role of the actors involved in the deed and that the contract is a legally binding agreement between the actors. In this way, formulaic expressions let us detect this structure.

3.2. Textual repetition across domains and genres

We compare the corpus of resolutions against a set of historical and modern text corpora from various domains, to get insight in how textually repetitive it is and whether that is related to the domain and genre of administrative and legal texts.

We use the following document collections to study the amount of textual repetition (Table 1). There are five corpora consisting of mostly administrative documents:

- The *Resolutions* corpus contains 286,871 printed resolutions of the SG in the period 1705-1796, with a Character Error Rate (CER) of 3%.
- The *Notarial Deeds of the city archive of Amsterdam* contain legal transactions. We expect deeds to have at least a few formulaic expressions stating the nature of the transaction, the parties involved, and that it has been signed off by a notary. We estimate the CER to be in around 10%.

[1] The latter aspect is part of the definition to account for historical variations in spelling of what is semantically or pragmatically the same phrase.

- The collection of *Dutch medieval charters* contains manual transcriptions of handwritten charters, which we expect to contain mostly formal language with low CER.
- The *Mandate and Police books from the city state of Bern* cover the same period (and more) and also contain administrative and legal documents where we expect formal language use, but in a different language (German) and with a higher rate of text recognition errors (CER of around 20%). The two types of books represent different administrative subgenres so we treat them as separate corpora.

We compare these against five corpora with mostly free-form text from various domains, including two with historic Dutch language and three corpora with modern Dutch:

- The *general missives of the Verenigde Oostindische Compagnie* (VOC, Dutch East India Company) consist of long business correspondences between the offices of the VOC in Asia and Amsterdam, which we expect are less formal and more free-form than the resolutions. These are handwritten documents, for which we could not obtain accurate CER information, but we estimate it to be in the range of 15-20%.
- The *Dutch Newspaper corpus* of the National Library of the Netherlands contains over 700,000 articles from dozens of Dutch language newspapers in the 18th century. This corpus was OCR'ed around 2006 and has a CER of 15-20%. The articles are free-form and cover many topics, so we expect low repetition.
- *Dutch Wikipedia* consisting of articles that are in principle free-form, but occasionally, entire databases of e.g. sports clubs, television shows or plant species are algorithmically turned into a set of article stubs using a template article. Although later manual edits of a template-based article tend to transform the template text to more free-form prose, some template phrasings may remain.
- *Dutch novels*, a set of 10,921 recently published Dutch novels (text extracted from epubs). We expect these to be free-form with little repetition.
- *Book reviews* is a set of 472,810 online Dutch book reviews [9, 21] from seven different reviewing platforms. Reviews are also free-form, although book reviews represent a very narrow domain with potentially many stock phrases, so we expect some form of repetition.

A *Vocabulary Growth Curve* [3, 6] shows how much the frequency of vocabulary terms grows with respect to the fraction of terms in a collection that have been seen only once. By iterating over all paragraphs in a corpus in a random order, we count the total number of terms N seen so far (i.e. term tokens) and at fixed points—e.g. once every 1000 words—divide the size of the vocabulary $V(N)$ over the number of hapax legomena $V_1(N)$, e.g. terms that have occurred only once so far. A higher value of $\frac{V(N)}{V_1(N)}$ means more of the term frequency mass is taken by terms that occur more than once.

The vocabulary growth curves are shown in Figure 2 for frequencies of word n-grams with $n \in [1, 3, 5]$. The administrative corpora are shown with dashed lines, while the more free-form corpora are shown with dotted lines. Further, corpora with high CER are shown with dense lines (little horizontal space between the symbols), and the rest with more widely separated symbols.

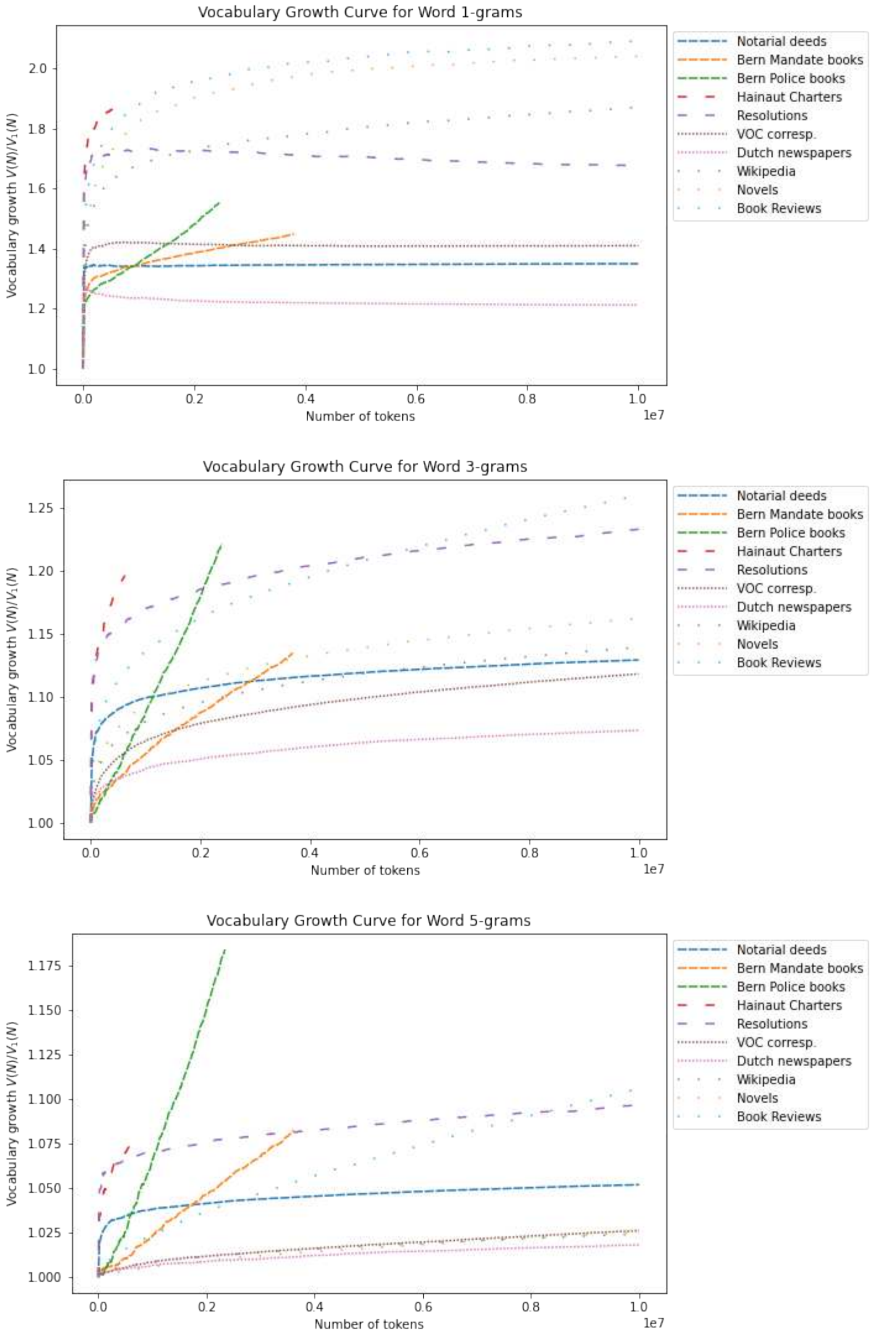


Figure 2: Vocabulary Growth curves of word n-grams for seven corpora of Dutch text. The top plot shows word 1-grams, the middle shows word 3-grams and the bottom shows word 5-grams. The corpora we assume to have formal language are represented by dashed lines, the free-form ones by dotted lines. Corpora with high Character Error Rate (CER) are represented by narrowly separated symbols, those with low CER by widely separated symbols.

Table 1
Overview of document collections used to study textual repetition

Collection name	Domain	Period	# docs	# words
Historic				
Resolutions of the Dutch States General	Administrative	1705-1796	286,871	58,430,762
Dutch East India Company missives	Administrative	1637-1792	981,457	218,923,640
Dutch Notarial deeds (Amsterdam)	Legal	1612-1833	93,262	29,298,606
Dutch medieval charters	Administrative	1299-1345	3,522	639,804
Dutch newspapers	News	1700-1799	705837	381,655,444
Bern region Mandate and Police books	Administrative	1458-1798	21,820	6,257,187
Modern				
Wikipedia NL	Reference	2005-2022	2,881,669	319,609,408
Dutch Novels	Fiction	2010-2020	10,921	745,977,872
Online Dutch book reviews	Reviews	1999-2020	472,810	57,970,421

For single words, the modern Dutch corpora have higher curves, meaning they have relatively more terms that occur more frequently, than the historic corpora based on algorithmic text recognition. This is not surprising, given the spelling variation and recognition errors in historic corpora. Both phenomena increase the size of the vocabulary and thereby lead to more mass at to the hapax legomena. Novels and reviews have the highest curves. We speculate that novels tend to use mostly common vocabulary to be easy to read by a large audience, while book reviews are a specific domain and genre, so use a relatively narrow vocabulary. Wikipedia contains articles about a huge range of topics, so it is understandable that it has a longer tail of hapax legomena. The medieval charters use a very limited vocabulary and have a lot of term repetition. As it is based on manual transcription, we assume the rate of recognition errors to be much lower than for the corpora based on OCR or HTR. The low/high CER distinction corresponds to a clear difference, with all low CER corpora having much higher curves.

For word 3-grams, the resolutions and Bern police and mandate books have curves that fall off very little, signalling that, although they have many single term hapax legomena, they have relatively many word 3-grams that occur more than once, compared to the other corpora. The curve for online book reviews overtakes the resolutions curve after around 500,000 tokens, suggesting indeed that additional reviews introduce relatively few new 3-grams and that reviews are therefore relatively similar to each other. For word 5-grams, the top curves are those of the charters, Bern police and mandate books, the resolutions, notarial deeds and the book reviews. With the exception of the latter, they are all in the domains of legal and administrative texts.

These curves support our intuition that texts in the legal domain have relatively many repeated phrases, despite the spelling variation and character recognition mistakes.

4. Modelling Formulaic Expressions

How frequent should a phrase be to be considered a formulaic expression? There can be formulaic expressions that are borrowed from other domains or genres, that are used with low frequency in the collection in which formulaic expressions are analysed. We leave these out of the scope of this paper, as we want to focus on expressions that are frequent enough that they can be used as metadata that cover most of the resolutions. Identifying borrowed formulas requires knowledge or analysis of external resources.

We want to find word sequences that occur frequently. The simplest way would be to count the frequencies of word n-grams for some range of n, similar to the word n-gram analysis of Section 3.2. However, the number of n-gram types grows rapidly as n increases, and the vast majority of these occur only once or twice. The corpus of resolutions has 735,919 distinct words, so the number of word 1-gram types is the same, but for $n = 4$, the number of n-gram types is 23,985,191. We exploit the fact that frequently occurring phrases can only consist of words that individually occur at least as frequently as the phrases themselves. That is, phrases that occur 100 times in the collection must consist of words that occur at least 100 times. To find phrases that occur at least $phrasefreq_{min} = 100$, we can exclude all candidate phrases that contain words with a corpus frequency below this phrase frequency threshold. Furthermore, the words within a phrase also co-occur with each other at least 100 times within a window that is equal to the word length of the phrase.

With these observations in mind, we developed a naive algorithm for identifying candidate formulaic expressions. In the pre-processing phrase, candidate phrases of fixed length are detected, after which their contexts of preceding and following words are clustered and analysed to extend the partial formulas and identify their start and end boundaries. The parameterisation we arrive at is ad hoc and specific this the corpus of resolutions, since we have no precise definition yet of what makes a phrase formulaic in a particular context. The goal is to explore the corpus with a pre-theoretical notion of what we are looking for.

Concretely, the pre-processing phase consist of the following steps:

1. Tokenise each resolution into sentences and sentences into words.
2. Iterate over the corpus and count frequencies of individual words
3. Iterate over the corpus a second time, and replace each word with a variable token <VAR> if it either has 1) a term frequency $termfreq(w_i) < phrasefreq_{min}$, or 2) a co-occurrence frequency $coocfreq(w_i, w_j) < phrasefreq_{min}$ with at least one of its remaining neighbouring $N = 5$ words $w_j \in (w_{i-N}, w_{i+N})$ on either side.
4. Slide a 5-word window over each individual sentence, extract the 5-word window as a phrase if it contains no variable token, and count the frequency of each extracted phrase.

The process is demonstrated in detail in Appendix A.

In the extension phase, we reduce the set of candidate phrases to a set of formulas in two steps. First, we use fuzzy string matching to find clusters of candidate phrases that are spelling variations of each other. In the second step, we gather the contexts around each occurrence of a cluster of phrases, and count how often the phrases are preceded and followed by the same sequence of words.

4.1. Clustering phrase spelling variants

Many of the common word 5-grams are spelling variants of each other. We cluster them by indexing these word 5-grams as vectors of character 1-skip-2-grams. That is, we consider not only 2 adjacent characters, but also pairs of characters that are separated by another character.

Starting from the most frequent word 5-gram phrases, we query the index to find candidate variants using cosine similarity. Further details are provided in Appendix B.

4.2. Extending partial formulas

Next, we build frequency lists of the 8 words preceding and following the fixed length phrase and use transition probabilities to identify extensions that have a probability close to 1 of preceding or following the phrase. This is inspired by probabilistic language models based on Hidden Markov Models [30, 17]. We note that there might be formulas shorter than 5 words. These can be detected by starting with shorter fixed length phrases.

Because the preceding (prefix) and following (postfix) contexts can include clusters of spelling variants as well, we use the same fuzzy matching algorithm as used for clustering the phrases. We then split all 8-word contexts into sequences of words and calculate transition probabilities for prefix and postfix contexts separately, starting from the fixed length phrase to the word immediately preceding or following it, and from that word to the next word, etc. Words that occur in multiple prefix or postfix contexts thereby have a higher transition probability. Words that have a probability below 0.1 are considered to be not part of the formulaic expression and are replaced by a <VAR> token. Once all transition probabilities have been computed, we traverse the transition model starting from the fixed length phrase and consider preceding words part of the formulaic expression if the probability is above 0.9. Once the probability drops below 0.9 but is still above 0.1, we assume to have reached a common context of the formula that is not part of the formula itself. We repeat the same process for the post-phrase context, again, computing transition probabilities starting from the fixed phrase.

This process is described in more detail in Appendix B.

5. Results

We start with the 23,141 candidate phrases that we got from using a minimum frequency threshold of 100. Clustering variant phrases reduces this to 12,880 clusters. The most frequent phrase is considered the representative variant. Extending these phrases with preceding and following words with a transition probability above 0.9 results in 11,497 candidate formulas and 52,348 common extensions (with cumulative transition probabilities $0.1 \leq p_{trans} < 0.9$).

The 10 most frequent phrases are shown in Table 2, together with their corpus frequencies and the formulas that were derived from analysing their contexts. The phrase '<START> ontfangen een missive van' is the most frequent fixed-length phrase that is also the most frequent formula (there is no extension that has a transition probability above 0.9). A less frequent and partially overlapping phrase is 'ontfangen een missive van den', which in the extension step is extended to '<START> ontfangen een missive van den', that is, the '<START>' token is added to it. The first two formulas therefore also partially overlap, but the second is an extension of the first. The word 'den' (EN: *the*) is the most common continuation of the first formula, but other common continuations are names of persons, so the second formula is less 'formulaic' than the first. Phrases 3 and 4 lead to the exact same formula, as do phrases 5, 7, 8 and 9. Phrase 10 is also a partial overlap with these phrases, but because it includes the word 'dat' (EN: *that*)

Table 2
The 10 most frequent word 5-gram phrases, their frequency and their associated formulas that were identified based on their contexts.

Rank	Initial phrase	Freq.	Formula
1	<START> ontfangen een missive van	138682	<START> ontfangen een missive van
2	ontfangen een missive van den	107679	<START> ontfangen een missive van den
3	geen resolutie is gevallen <END>	90265	waar op geen resolutie is gevallen <END>
4	op geen resolutie is gevallen	89528	waar op geen resolutie is gevallen <END>
5	waar op gedelibereert zynde is	86621	waar op gedelibereert zynde is goedgevonden en verstaan
6	waar op geen resolutie is	68153	waar op geen resolutie is gevallen <END>
7	op gedelibereert zynde is goedgevonden	67601	waar op gedelibereert zynde is goedgevonden en verstaan
8	zynde is goedgevonden en verstaan	66061	waar op gedelibereert zynde is goedgevonden en verstaan
9	gedelibereert zynde is goedgevonden en	64207	waar op gedelibereert zynde is goedgevonden en verstaan
10	is goedgevonden en verstaan dat	55974	zynde is goedgevonden en verstaan dat

at the end — which is again a common follow up, but not the only one — it is not extended to the same formula.

We can further cluster these formulas by re-categorising the less frequent formulas that are extended or reduced versions of more frequent formulas as common *extensions*. If we perform this re-categorisation, the list of 11,497 formulas is reduced 7,153 formulas. Eyeballing the list of remaining formulas reveals there are still many spelling variants in the list. This shows that spelling variant is a challenging problem that needs to be analysed in more detail.

5.1. Analysis of formulas

Almost 58% of the candidate formulas are longer than 5 words. Of course, our choice to start from candidate phrases of 5 words ensures no formulas shorter than 5 words are found, but the fact that more than half are extended shows that formulaic expressions in the resolutions are long syntactic units consisting of more than compound nouns and noun phrases.

The 10 formulas resulting from clustering the most frequent formulas are shown in Table 3. The last column describes how we can use these formulas to identify meaningful elements in the running text, and how they help in classifying these elements. It is worth noticing that many formulas precede of follow a named entity, suggesting that formulas were frequently used to assert the relationship of that entity to the proposition or decision. Several formulas also contain verbs (*has been read*, *has been agreed and understood*, *to report back*) that signal specific actions. In future work, we will analyse the types of relationships and actions that these formulas express.

Table 3
The top 10 most frequent formulas and how they structure and identify elements of the text.

Rank	Formula	Translation	Signal
1	<START> ontfangen een missive van	<START> received a missive of	this is the start of a resolution and the start of the proposal paragraph, the proposal document type is *missive*
2	waar op geen resolutie is gevallen <END>	on which no resolution was made <END>	this is the end of the resolution, no decision was made
3	waar op gedelibereert zynde is goetgevonden en verstaan	which, upon deliberation, has been agreed and understood	start of the decision paragraph
4	en van alles alhier ter vergaderinge rapport te doen <END>	and to report back on everything, here in the meeting	decision to start a investigative committee that will report back at a later date, signal that there is a future resolution related to this one
5	en andere haar hoogh mogende gedeputeerden tot de	and other high and mighty deputies of	name preceding this formula is a deputy, name following this formula is an institution, the deputy is a representative of the institution
6	de heeren gedeputeerden van de	the gentlemen deputies of the	the name following this formula is a province or institution
7	in handen van de heeren	in the hands of the gentlemen	decision that the matter is handed to a committee to investigate, the name(s) following this formula are the members of this committee
8	haar hoogh mogende resolutie van den	resolution of her high and mighty of the	what follows is a date of a previous resolution, the previous resolution is related to this resolution
9	aan het hof van sijne	at the court of his	the name preceding this formula is a representative of the court of the name following this formula
10	Is ter vergaderinge gelesen de requeste van	Has been read in this meeting, the petition of	this is the start of a resolution and the start of the proposal paragraph, the proposal document type is *missive*

Applying the approach to other corpora is elaborated in Appendix E.

5.2. Challenges of Evaluation

The detection approach above is exploratory, as we have no precise definition of what a formulaic expression is. We need clear criteria to determine if a phrase is formulaic before we can precisely define the task of formula detection and quantitatively evaluate methods designed to perform that task. For a proper evaluation of the detection method the entire corpus needs to be annotated with all formulas. We could reduce the problem by focusing on a small sample of resolutions and annotate anything that we think is a formula, but we would still need clear criteria to decide what is a formula and what is not. Another alternative is to use simulation and generate text and insert artificially generated formulaic expressions to fully control the characteristics of formula, including their length, variation and frequency of occurrence.

Intuitively, a quantitative evaluation should consider precision and recall of detecting formulaic expressions, with two different measures of recall. One is the fraction of different types of formulaic expressions that are identified (type-based recall), and the other is the fraction of occurrences of formulaic expressions that are detected (token-based recall).

6. Discussion and Conclusions

Formulas and their use have relevance to a number of both methodological and humanities research questions. Because formulaic expressions and their use for text structuring are understudied, for a large part we can only raise these research questions.

Our findings of the use of repetitive phrases in various corpora suggests that repetitive language use differs strongly across domains and genres, with texts in administrative domains containing more repetitive phrasing. It is not clear whether this is true for all or just a specific type of administrative texts. Our findings with detecting formulas suggest that this type of serial government sources may contain many formulaic expressions that can be exploited to structure texts and extract information. Further research has to shed light on the extent to which this also holds for other administrative sources and how the composition and diversity of collections relates to the statistical properties of formulas. A second type of research questions centres on the use of formulas. We do not know whether comparable administrative archives from different periods have the same rate of repetitive phrases and formulas. We observed that formulas emerged, changed and disappeared over time, but not at what rate and what the causes were. We do not know if there was an increase in adoption of formulaic expressions in the resolutions. There could have been changes in legal or administrative customs, more general language and cultural changes or perhaps even influences of specific scribes. The switch to the use of printing would lead us to assume that there was less variation in the formulas used, but it is too early to test this assumption. It is also an open question where the formulas originate, and whether formulas are reused across (administrative) domains. A formula borrowed from another domain might not be used often in a corpus, in which case the method and definition we developed in this paper do not suffice. So further discussion is needed on what constitutes a formulaic expression, and what its generic and context-specific elements are. As for the content of formulas, we can only make some remarks about those that occur in the resolutions. We have been able to spot a great number of formulas but it is hard to give a precise definition of what a formula is, or to establish if we can discern constituent elements and if it makes sense to divide up formulas in these constituent parts. In this paper we describe a methodology in development, that iteratively gathers formulas from our corpus. This works well for identifying the most frequently used formulas and their variations. Further steps have to make clear what the optimal rate of formula detection is, how to categorise the different formulas, and if they are sufficient, to identify and localise different logical elements in the text and whether this works for all resolutions. A further point of discussion is how much of the methodology can be used for other corpora. We believe that the general methodology is suitable for comparable text corpora.

7. Acknowledgments

This work is part of the REPUBLIC project (2019-2024), a Research Infrastructure project funded by the Dutch Research Council (NWO, Grant number 175.217.024).

References

[1] G. Altmann and R. Köhler. "Forms and Degrees of Repetition in Texts". In: *Forms and Degrees of Repetition in Texts*. De Gruyter Mouton, 2015.

[2] P. Auslander. "On Repetition". In: *Performance Research* 23.4-5 (2018), pp. 88–90.

[3] H. Baayen. "The effects of lexical specialization on the growth curve of the vocabulary". In: *Computational Linguistics* 22.4 (1996), pp. 455–480.

[4] E. R. Babbie. *The practice of social research*. Cengage learning, 2020.

[5] A. Baron and P. Rayson. "VARD2: A tool for dealing with spelling variation in historical corpora". In: *Postgraduate conference in corpus linguistics*. 2008.

[6] M. Baroni and S. Evert. "The zipfR package for lexical statistics: A tutorial introduction". In: *Available atzipfr. r-forge. r-project. org/materials/zipfrtutorial. pdf [last accessed1 June 2019]* (2014).

[7] D. Biber, S. Johansson, G. Leech, S. Conrad, E. Finegan, and R. Quirk. *Longman grammar of spoken and written English*. Vol. 2. Longman London, 1999.

[8] B. Boguraev and C. Kennedy. "Technical Terminology for Domain Specification and Content Characterisation". In: *Scie*. 1997. DOI: 10.1007/3-540-63438-x_5.

[9] P. Boot. "A Database of Online Book Response and the Nature of the Literary Thriller." In: *Dh*. 2017.

[10] R. L. Cannon. "OPCOL: An Optimal Text Collation Algorithm". In: *Computers and the Humanities* (1976), pp. 33–40.

[11] D. S. Carvalho, M.-T. Nguyen, C.-X. Tran, and M.-L. Nguyen. "Lexical-Morphological Modeling for Legal Text Analysis". In: *JSAI International Symposium on Artificial Intelligence*. Springer, 2015, pp. 295–311.

[12] D. S. Carvalho, V. D. Tran, K. Van Tran, V. D. Lai, and M.-L. Nguyen. “Lexical to Discourse-Level Corpus Modeling for Legal Question Answering”. In: *Tenth International Workshop on Juris-Informatics (JURISIN).* 2016.

[13] K. Conklin and N. Schmitt. “The processing of formulaic language”. In: *Annual Review of Applied Linguistics* 32 (2012), pp. 45–61.

[14] A. P. Cowie. *Phraseology: Theory, analysis, and applications.* OUP Oxford, 1998.

[15] R. Haentjens Dekker, D. Van Hulle, G. Middell, V. Neyt, and J. Van Zundert. “Computer-supported collation of modern manuscripts: CollateX and the Beckett Digital Manuscript Project”. In: *Digital Scholarship in the Humanities* 30.3 (2015), pp. 452–470.

[16] I. Hendrickx and R. Marquilhas. “From Old Texts to Modern Spellings: An Experiment in Automatic Normalisation.” In: *J. Lang. Technol. Comput. Linguistics* 26.2 (2011), pp. 65–76.

[17] F. Jelinek. *Statistical methods for speech recognition.* MIT press, 1998.

[18] A. Joulin, E. Grave, P. Bojanowski, and T. Mikolov. “Bag of tricks for efficient text classification”. In: *arXiv preprint arXiv:1607.01759* (2016).

[19] D. Jurafsky and J. H. Martin. “Speech and Language Processing: An introduction to speech recognition, computational linguistics and natural language processing”. In: *Upper Saddle River, NJ: Prentice Hall* (2008).

[20] F. Karsdorp. “Het is groen en leeft nog lang en gelukkig. Classificatie van volksverhaalgenres op basis van formules”. In: *Tijdschrift voor Nederlandse Taal-en Letterkunde* 129.4 (2014), pp. 274–288.

[21] M. Koolen, P. Boot, and J. J. van Zundert. “Online Book Reviews and the Computational Modelling of Reading Impact”. In: *Proceedings of the Workshop on Computational Humanities Research (CHR 2020).* Vol. 2723. 2020, p. 0073. URL: http://ceur-ws.org/Vol-2723/long13.pdf.

[22] M. Koolen, R. Hoekstra, I. Nijenhuis, R. Sluijter, E. van Gelder, R. van Koert, G. Brouwer, and H. Brugman. “Modelling Resolutions of the Dutch States General for Digital Historical Research.” In: *Colco.* 2020, pp. 37–50.

[23] E. Manjavacas, Á. Kádár, and M. Kestemont. “Improving Lemmatization of Non-Standard Languages with Joint Learning”. In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers).* Minneapolis, Minnesota: Association for Computational Linguistics, 2019, pp. 1493–1503. DOI: 10.18653/v1/N19-1153. URL: https://www.aclweb.org/anthology/N19-1153.

[24] V. T. Martins, D. Fonte, P. R. Henriques, and D. d. Cruz. “Plagiarism detection: A tool survey and comparison”. In: (2014).

[25] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean. “Distributed representations of words and phrases and their compositionality”. In: *Advances in neural information processing systems* 26 (2013).

[26] M. Poß and T. v. d. Wouden. "Extended lexical units in Dutch". In: *LOT Occasional Series* 4 (2005), pp. 187–202.

[27] M. Potthast, A. Eiselt, L. A. Barrón Cedeño, B. Stein, and P. Rosso. "Overview of the 3rd international competition on plagiarism detection". In: *CEUR workshop proceedings.* Vol. 1177. CEUR Workshop Proceedings. 2011.

[28] M. Potthast, M. Hagen, T. Gollub, M. Tippmann, J. Kiesel, P. Rosso, E. Stamatatos, and B. Stein. "Overview of the 5th international competition on plagiarism detection". In: *CLEF Conference on Multilingual and Multimodal Information Access Evaluation.* Celct. 2013, pp. 301–331.

[29] M. Potthast, B. Stein, E. Andreas, and A. B.-C. P. Rosso. "Overview of the 1st international competition on plagiarism detection". In: *3rd PAN Workshop. Uncovering Plagiarism, Authorship and Social Software Misuse.* 2009, p. 1.

[30] L. R. Rabiner. "A tutorial on hidden Markov models and selected applications in speech recognition". In: *Proceedings of the IEEE* 77.2 (1989), pp. 257–286.

[31] G. E. Raney. "A Context-Dependent Representation Model for Explaining Text Repetition Effects". In: *Psychonomic Bulletin & Review* 10.1 (2003), pp. 15–28.

[32] R. Rashidi-Tabrizi, G. Mussbacher, and D. Amyot. "Legal Requirements Analysis and Modeling with the Measured Compliance Profile for the Goal-Oriented Requirement Language". In: *2013 6th International Workshop on Requirements Engineering and Law (RELAW).* Ieee, 2013, pp. 53–56.

[33] R. Rehurek and P. Sojka. "Gensim–python framework for vector space modelling". In: *NLP Centre, Faculty of Informatics, Masaryk University, Brno, Czech Republic* 3.2 (2011), p. 2.

[34] M. Reynaert. "Non-interactive OCR post-correction for giga-scale digitization projects". In: *International Conference on Intelligent Text Processing and Computational Linguistics.* Springer. 2008, pp. 617–630.

[35] M. Reynaert. "TICCLops: Text-Induced Corpus Clean-up as online processing system". In: *Proceedings of coling 2014, the 25th international conference on computational linguistics: System demonstrations.* 2014, pp. 52–56.

[36] S. Ruecker, M. Radzikowska, P. Michura, C. Fiorentino, and T. Clement. "Visualizing Repetition in Text". In: *Digital Studies/Le champ numérique* 1.3 (2009).

[37] G. Rutten and M. J. van der Wal. "Functions of epistolary formulae in Dutch letters from the seventeenth and eighteenth centuries". In: *Journal of Historical Pragmatics* 13.2 (2012), pp. 173–201.

[38] H. Salmi, P. Paju, H. Rantala, A. Nivala, A. Vesanto, and F. Ginter. "The reuse of texts in Finnish newspapers and journals, 1771–1920: A digital humanities perspective". In: *Historical Methods: A Journal of Quantitative and Interdisciplinary History* 54.1 (2020), pp. 14–28.

[39] M. A. Samkova. "Repetition and Intertextuality as Modalities of Text Structuring and perception". In: *Facta Universitatis, Series: Linguistics and Literature* 0 (2016), pp. 95–105.

[40] D. Schmidt and R. Colomb. "A data structure for representing multi-version texts online". In: *International Journal of Human-Computer Studies* 67.6 (2009), pp. 497–514.

[41] N. Schmitt. *Formulaic sequences: Acquisition, processing, and use.* Vol. 9. John Benjamins Publishing, 2004.

[42] N. Schmitt and R. Carter. "Formulaic sequences in action". In: *Formulaic sequences: Acquisition, processing and use* (2004), pp. 1–22.

[43] E. Stamatatos. "Intrinsic plagiarism detection using character n-gram profiles". In: *threshold* 2.1,500 (2009).

[44] A. Vesanto, F. Ginter, H. Salmi, A. Nivala, and T. Salakoski. "A system for identifying and exploring text repetition in large historical document corpora". In: *Proceedings of the 21st Nordic Conference on Computational Linguistics.* 2017, pp. 330–333.

[45] A. Vesanto, A. Nivala, H. Rantala, T. Salakoski, H. Salmi, and F. Ginter. "Applying BLAST to text reuse detection in finnish newspapers and journals, 1771-1910". In: *Proceedings of the NoDaLiDa 2017 Workshop on Processing Historical Language.* 2017, pp. 54–58.

[46] D. Wood. "Uses and functions of formulaic sequences in second language speech: An exploration of the foundations of fluency". In: *Canadian Modern Language Review* 63.1 (2006), pp. 13–33.

[47] A. Wray. *Formulaic language and the lexicon.* Eric, 2002.

8. Appendix

A. Identifying candidate partial formulas

We demonstrate the processing steps to identify candidate partial formulas using the following example sentences:

> Ontfangen een Missive van het Collegie ter Admiraliteyt in Zeelandt, geschreven te Middelburgh den negentienden deser loopende maandt, houdende, in gevolge en tot voldoeninge van haar Hoogh Mogende Resolutie van den vyfden der voorlede maandt, der zelver advis op het verzoeck van Burgermeesters en Scheepenen van het hooge en laage Zas van Gent.

With word tokenisation, we add <START> and <END> tokens so that for words at the start or end of a sentence, this boundary is included in its context. Certain phrases only appear at or near the start or end of a sentence, and adding boundary tokens allows us to keep track of such cases.

After filtering out low frequency words and words that do not meet the co-occurrence frequency threshold, we end up with the following list:

> [‘<START>’, ’ontfangen’, ’een’, ’missive’, ’van’, ’het’, ’collegie’, ’ter’, ’admiraliteyt’, ’in’, ‘<VAR>’, ’geschreven’, ’te’, ‘<VAR>’, ’den’, ‘<VAR>’, ’deser’, ’loopende’, ’maandt’, ’houdende’, ’in’, ‘gevolge’, ’en’, ’tot’, ’voldoeninge’, ’van’, ’haar’, ’hoogh’, ’mogende’, ’resolutie’, ’van’, ’den’, ‘<VAR>’, ’der’, ’voorlede’, ’maandt’, ’der’, ’zelver’, ’advis’, ’op’, ’het’, ‘<VAR>’, ’van’, ‘<VAR>’, ’en’, ‘<VAR>’, ’van’, ’het’, ‘<VAR>’, ’en’, ‘<VAR>’, ‘<VAR>’, ’van’, ‘<VAR>’, ‘<END>’]

The example above shows that most of the words in the sentence occur frequently and co-occur with each other frequently. This results in a large number of candidate phrases. In the resulting sentence, we extract candidate phrases $p(w_i, w_{i+5}$ by identifying sequences of word tokens (i.e. non-variable tokens) of length $|p| = 5$ and count their frequency over the entire corpus.

Table 4
The impact of the minimum term and co-occurrence frequency thresholds on the vocabulary and number of co-occurring word pairs and number of phrases.

Min. Freq.			Total 5-word phrases		Phrases above
Threshold	Vocab. size	Co-oc pairs	Types	Tokens	Threshold
1	614,829	27,345,551	30,261,003	57,027,160	30,261,003
10	71,151	19,260,903	24,270,577	50,687,254	282,486
100	17,021	12,025,517	16,694,381	41,427,641	23,141
1000	3,547	3,993,786	6,839,379	26,668,452	2,119
10,000	605	316,686	756,629	10,435,346	140

As little is known in advance about the relationship between frequencies of words, word co-occurrence and formulas, we have few meaningful clues for setting a minimum phrase frequency. The corpus of resolutions has 286,871 resolutions and over 58 million words, but we have no reliable estimates of the total number of formulaic expressions that are used and how often they occur. In identifying the start of proposition paragraphs of resolutions, [22] use a list of 32 formulas, most of which are between 5 and 10 words long. But we do not know if these are all the formulas used for openings of proposition and decision paragraph. We also do not know which other elements of a resolution are expressed in formulas.

To get some insight, we experimented with minimum frequencies of different orders of magnitude, ranging from 1 to 10,000. The results are shown in Table 4, with for each frequency threshold, the size of the vocabulary (number of distinct words), the number of distinct co-occurring word-pairs within a 5 word window, the number of distinct 5-word phrases (phrase types) after pre-processing, the total number of phrases (phrase tokens) and the number of phrases that meet the frequency threshold.

Frequency thresholds of 1 and 10 lead to large vocabularies and huge numbers of phrases. If they are to be made useful as metadata labels or boundary signals, we need to analyse each of them in their context manually, so these thresholds result in an unmanageable number of phrases. On the other extreme, a threshold of 10,000 leads to a very small vocabulary of 605 highly common words and only 140 highly frequent phrases. The most common phrase is *<START> Ontfangen een Missive van'* (EN: *<START> Received a missive of*), which is also the start of the example sentence above, with a frequency of 138,682. Note that there are variant spellings of this phrase that are not included in the count. With such a high frequency, it is clear that this phrase is part of a fixed formula, but because we made fixed-length phrases, it is not clear whether this the entire formula.

It is quite likely that some formulas contain words with a frequency below 10,000, and because we know so little yet about the usage of formulas, it is better to be conservative and choose a lower threshold. For the rest of this paper, we pick a minimum frequency threshold $phrasefreq_{min} = 100$. This gives us a large set of 23,141 frequent phrases that are candidate formulaic expression or part of them.

B. From candidate phrases to formulaic expressions

The process of transforming the set of candidate phrases to a set of formulas has two main steps. First, we use fuzzy string matching to find clusters of candidate phrases that are spelling variations of each other. In the second step, we gather the contexts around each occurrence of a cluster of phrases, and count how often the phrases are preceded and followed by the same sequence of words.

B.1. Clustering variant phrases

We index each word 5-gram phrase as a vector of character 1-skip-2-grams. That is, we consider not only 2 adjacent characters, but also pairs of characters that are separated by another character. The reason to include a single skip is that spelling variants often have differences in characters in the middle of a word, which results in multiple ngram mismatches and thus few matches when no skips are used.

Starting from the most frequent word 5-gram phrases, we query the index to find candidate variants using cosine similarity. Further details are provided in Appendix B. We limit the candidate set to phrases that differ in length with the query phrase by at most 2 characters, based on the assumption that much longer or shorter phrases are unlikely to be spelling variants. We filter the candidate phrases by checking that the words in each position 1..5 of the candidate phrase differ in length no more than 2 characters with their aligned words in the query phrase. This avoids matching a query phrase 'op gedelibereert zynde is goetgevonden' with its partial overlap 'gedelibereert zynde is goetgevonden en'. The latter is typically the next 5-word phrase following the former, but because the query phrase has a short first word and the candidate phrase has a short last word, their ngram similarity is high. Using the length restriction on aligned words filters out such erroneous matches.

B.2. Extending candidate formulas

In the extension step, we build frequency lists of the 8 words preceding and following the fixed length phrase and use transition probabilities to identify extensions that have a probability close to 1 of preceding or following the phrase. This is a similar approach to probabilistic language modelling [19], where the probably that a word w_i is followed by word w_j, is calculated as:

$$P(w_j|w_i) = \frac{P(w_i, w_j)}{P(w_i)} \tag{1}$$

This models the prediction of the next word only the current word. A natural extension is to model the prediction based on all words preceding it:

$$P(w_j|w_1, w_2, ..., w_i) = \frac{P(w_1, w_2, ..., w_i, w_j)}{P(w_1, w_2, ..., w_i)} \tag{2}$$

For extending phrases, we use this model, where the probability $P(phr)$ of a phrase phr consisting of words $w_i, ..., w_j$ is $P(w_i, ..., w_j)$. The transition probability from a phrase phr to a word w_i is $\frac{P(phr,w_i)}{P(phr)}$ and from a word w_i to phr is $\frac{P(w_i,phr)}{P(phr)}$.

We split all 8-word contexts into sequences of words and calculate transition probabilities for prefix and postfix contexts separately, starting from the fixed length phrase to the word immediately preceding or following it, and from that word to the next word, etc. Words that occur in multiple prefix or postfix contexts thereby have a higher transition probability. Words that have a probability below 0.1 are considered to be not part of the formulaic expression and are replaced by a <VAR> token. Once all transition probabilities have been computed, we traverse the transition model starting from the fixed length phrase and consider preceding words part of the formulaic expression if the probability is above 0.9. Once the probability drops below 0.9 but is still above 0.1, we assume to have reached a common context of the formula that is not part of the formula itself. We repeat the same process for the post-phrase context, again, computing transition probabilities starting from the fixed phrase.

We extend the fixed length phrase with up to 8 words preceding it and up to 8 words following it, which means we can identify formulas of $8 + 5 + 8 = 21$ words in a single pass. If the full 8-word path preceding or following the phrase has a cumulative transition probability close to 1, we repeat this extension process with the extended phrase to identify the boundary of the formula.

An example of extending the phrase 'gevolge en tot voldoeninge van' using transition probabilities is shown in Figure 3. On the left side, the prefix context is shown, with the word 'in' being the only word directly preceding the partial phrase. This means that 'gevolge en tot voldoeninge van' is not the full formulaic expression, but that 'in' is also part of it. The expression 'in gevolge en tot voldoeninge van' is also a syntactically more comprehensible phrase, meaning *in consequence and fulfilment of.* There are multiple possible words preceding it. Two common ones are the verbs 'hebbende' (EN: *having*), which precedes the phrase 'in gevolge en tot voldoeninge van' in 38% of the occurrences of the phrase, and 'houdende' (EN: *maintaining*) which precedes the phrase in 47% of its occurrences. The remaining occurrences of the phrase are preceded by a variety of other words. Each of these two verbs have multiple possible prefixes, but are themselves not part of the formula according to the definition above, because their transition cumulative transition probabilities ($1.0*0.47 = 0.47$ and $1.0*0.38 = 0.38$ respectively) do not meet the threshold of 0.9.

On the right side, the postfix context is shown, with two variants of continuations, neither of which is part of the formulaic expression itself. One common continuation is 'der selver resolutie' and the other is 'haar hoog mogende resolutie'. Note that although neither path to the word 'resolutie' has itself a cumulative transition probability above 0.9 ($0.39 * 0.99 * 0.99 = 0.38$ for the former and $0.59 * 0.99 * 0.99 * 0.96 = 0.56$ for the latter), their combined probabilities add up to 0.94. This meets the 0.9 probability threshold, thereby being an example of a formulaic expression that can have a variable middle part. In some cases that variable part is 'der selver' and in others it is 'haar hoogh mogende'.

The final formula is determined by extending the fixed length phrase with preceding and following words that have a cumulative probability of at least 0.9. In the case of the phrase 'gevolge en tot voldoeninge van', we extend with the prefix 'in' to 'in gevolge en tot voldoeninge van'

C. The impact of spelling change

One of the big hurdles is spelling change. Finding phrases that are orthographically similar to each other is not hard, but phrases often contain highly frequent, short function words that require only few character edits to transform one function word into another. This makes it difficult to distinguish cases where two function words are variant spellings of each other from cases where they represent different words and therefore signal that these phrases have different meanings.

We experimented with both classic Word2Vec [25] and fastText [18] CBOW embeddings[2] to identify variant spellings of words, choosing the latter as it has better performance on a test set of target words and their variants. FastText uses character-level embeddings that are more

[2] For both models we use their implementations in Gensim 4.0 [33], see https://radimrehurek.com/gensim/

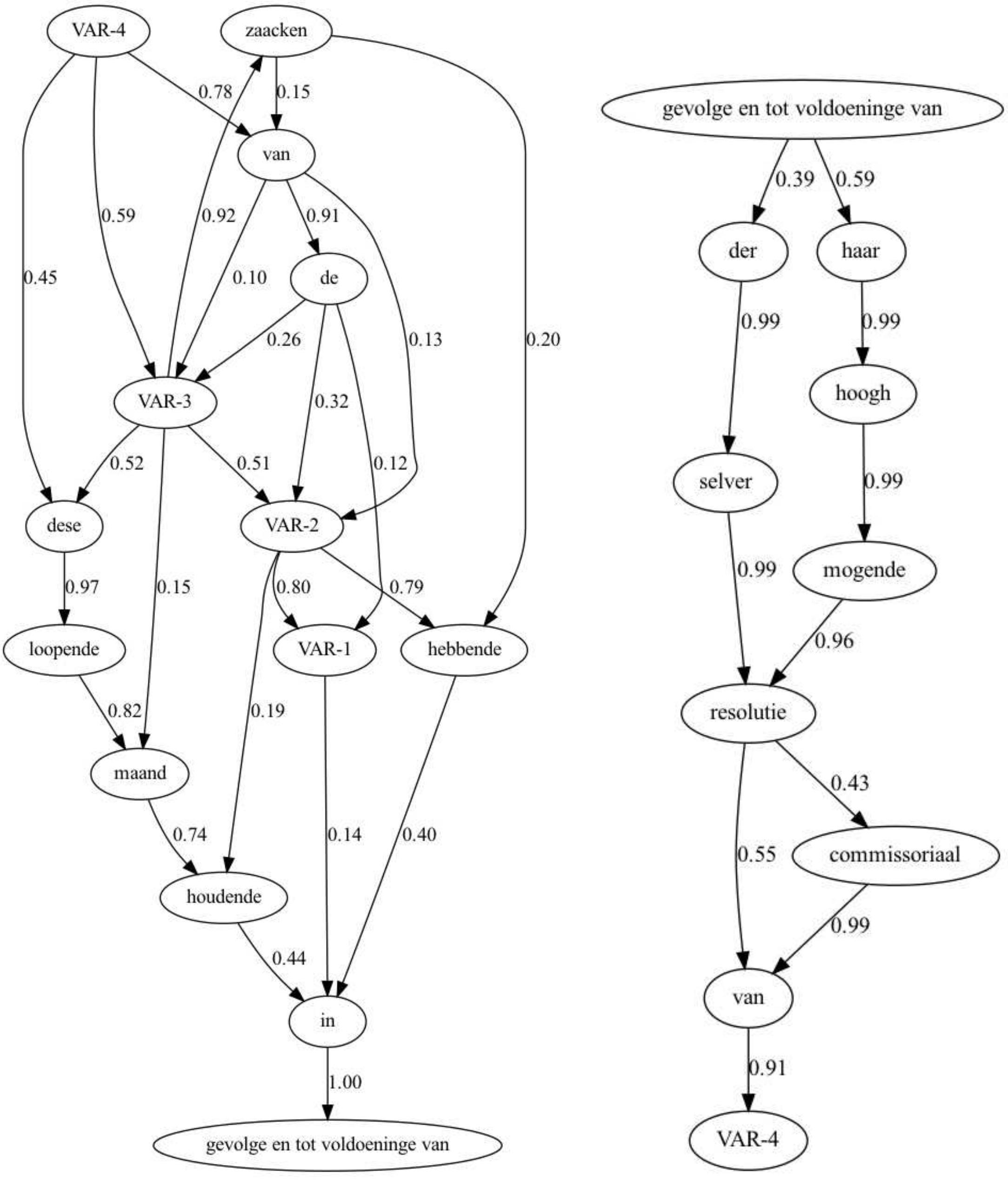


Figure 3: Extending the phrase 'gevolge en tot voldoeninge van' using transition probabilities. The prefix probabilities are on the left, the postfix probabilities on the right.

suitable for detecting spelling variations. We use embeddings based on the assumption that variants occur in the same or similar contexts so should end up in the same region in the embedding space. This works for spelling variants that are used interchangeably in a single time window. An example in the Resolutions corpus is the word 'en' (EN: *and*) and its variant 'ende'.

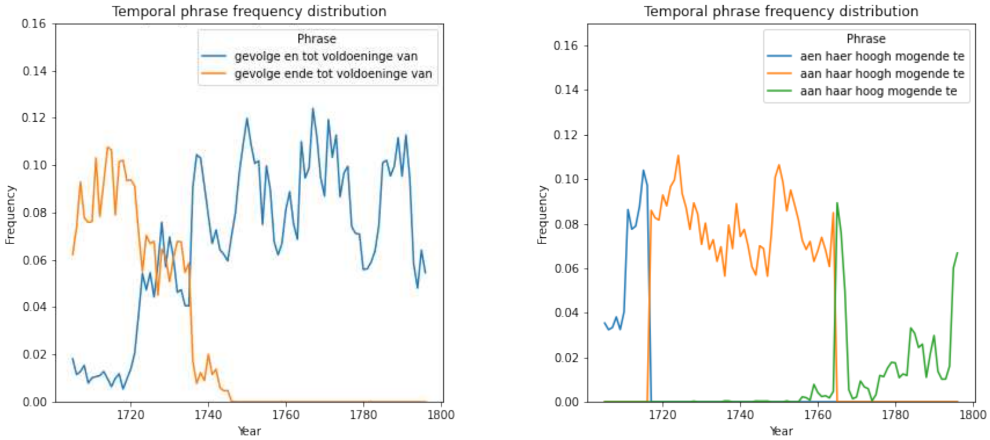


Figure 4: The distribution of yearly frequencies of partial phrases in the 18th century resolutions.

There is a period in the 18th century when their uses overlap, as can be seen in the temporal frequency distributions of the variant phrases ’gevolge en tot voldoeninge van’ (EN: *following and in fulfilment of*) and ’gevolge ende tot voldoeninge van’ (see the left side of Figure 4). The two versions ‘en’ and ‘ende’ share many contexts, so their word embeddings are similar. Using a combination of orthographic similarity (edit distance) and embedding similarity, we can identify word pairs in the corpus that can be linked as variants. However, experiments on finding variant spellings of words in the pre- and post-context of phrases have shown that for short function words, spelling change is a major hurdle for both Word2Vec and FastText. But for spelling changes where one variant is used in only one period and the other only in another, non-overlapping period, their contexts can also have different spellings. This results in two sets of contexts that also have no or little overlap. Hence, two spelling variants used in different time periods may end up in different regions in the embedding space. An example is the use ’ae’ in the early 18th century in words like ’aen’ (EN: *to* and ’haer’ (EN: *her*), which changed to using ’aa’ from around 1717, when they switched to writing ’aan’ en ’haar’. These words often appear together, as in the common phrase ’aen haer hoogh mogende te’ (EN : *to her high and mighty at*). Because the spelling change for ’aen’ occurs at the same time as the spelling change for its common contextual term ’haer’, the variants ’aen’ and ’aan’ have little contextual overlap, so word embeddings consider them as different words.

D. The impact of resolution length

One of the characteristic of resolutions that we can study with our list of formulas is the fraction of a resolutions text is made up of formulaic expressions, and which part is not. There are many very short resolutions based on a received missive (starting with the formula ‘Ontfangen een Missive van’) that merely states who wrote the missive, when and where they wrote it, but that do not contain any proposal or request that the SG had to make a decision on. Such resolutions

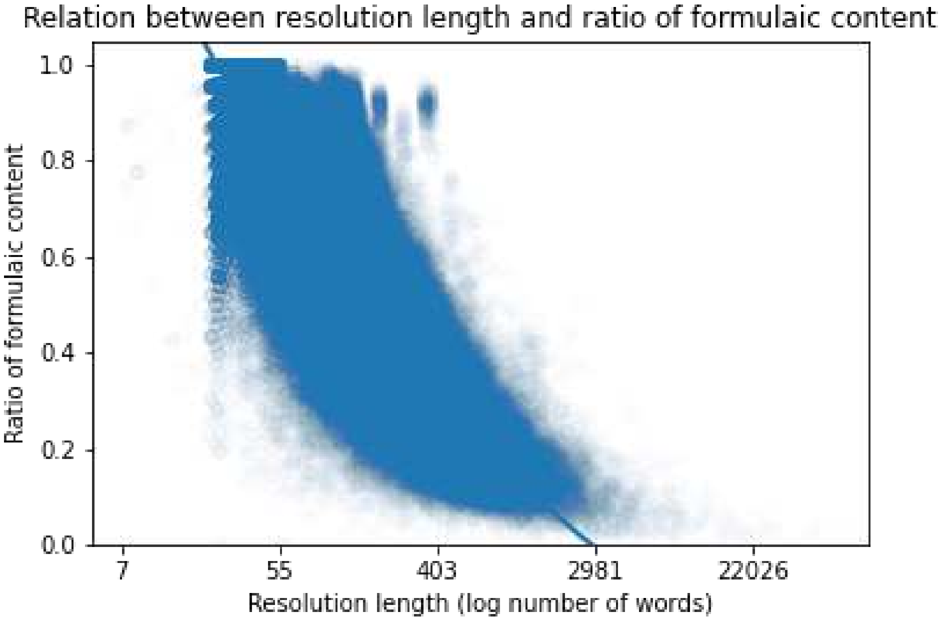


Figure 5: The relationship between the length of a resolution and the fraction of words in a resolution that are part of a formula.

end with the formula 'Waar op geen resolutie is gevallen.' (EN: *On which no decision was made.*

We therefore expect there to be a relationship between the length of a resolution and the amount of non-formulaic content. Longer resolution provide more detail of the proposition or of the decision or both. We assume that these details are only given when they are deemed relevant and necessary. The details vary across resolutions, therefore lead to less formulaic text.

The relationship between the length of resolutions and the fraction of words that are part of formulaic expressions is shown in Figure 5. There is a clear relationship: short resolutions tend to have a larger fraction of formulaic content. As resolutions get longer, a larger fraction of the words they contain are below one of the two frequency thresholds.

E. Formulas in Other Corpora

To check if this approach generalises to other corpora, we use the same detection process on the corpora of Notarial Deeds, Bern Manate books and t he Dutch Wikipedia.

E.1. Mandate books of State of Bern

Because of the high Character Error Rate (CER∼ 0.2) and the much smaller size of the corpus (3.8 million words compared to 58 million of the Resolutions), we used word 4-grams instead of 5-grams. The procedure found a handful of candidates, two of which could be extended to formulaic expressions:

- The most common phrase is ’Schultheiß und Rath der Statt Bern’, which occur 907 times in 505 different spellings. It refers to the head official and the council of the city of Bern.
- The second most frequently identified phrase is ’An alle Deütsch und Weltsche’, which occurs 559 times in 443 different spellings, and is extended to the formula ‘An alle Deütsch und Weltsche Herren Amtleüth’ (EN: *To all German and X gentlemen officials*). This is a formula to signal that the following statute pertains to the officials of both the French and German speaking parts of the city state Bern, and therefore signals the start of a statute.

E.2. Notarial deeds from Amsterdam municipality

The notarial deeds corpus also has a high CER of 15-20%. The most frequently found phrase are:

- ‘als getuijgen hier overgestaen’ (EN: *standing here as witnesses*: this is part of a formulaic phrase in the opening paragraph of a notarial deed to indicate who act as witnesses in formalising the transaction. This is useful in identifying the starting paragraphs of deeds that are spread across pages.
- ‘H Schaef N P’: the name one of the Amsterdam notaries, who is the official responsible for ensuring the transaction is legal.
- ‘J de Winter N P’: the name of another Amsterdam notary.

These phrases have similar potential in identifying meaningful structural elements in the running text, such as where deeds start and end, and where certain elements of deeds are located within their text.