The Location and Function of Formulaic Expressions in the Resolutions of the Dutch States General

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

[1]Huygens Institute
[2]DHLab - KNAW Humanities Cluster

1 Introduction

Formulaic expressions are commonly used in administrative documents to signal important aspects of a document (Koolen and Hoekstra, 2022, Kopaczyk, 2012, 2013). Medieval charters contain opening and closing formulas to signal that the document is a charter and what type of charter it is (Boonen, 2005, De Boor, 1975). Notarial deeds contain formulas based on notary manuals to make sure the transaction they confirm is unambiguous and follows protocol (Lemercier and Trivellato, 2022, Marques, 2018, Zomeño, 2007).

In previous work, we developed techniques to automatically detect formulas in historic document collections, while dealing with orthographic variation introduced by historic spelling variation and change and errors introduced by OCR and HTR processes Koolen and Hoekstra (2022). In this paper, we investigate the nature of the formulas detected in the resolutions of the Dutch States General.

The resolutions are transcripts of the decisions of the Dutch States General (SG) in the period 1576-1796, estimated one million in total. Each resolution consists of at least two parts, a proposition and a decision (Thomassen, 2019). Most resolutions contain no more than this, but some resolutions have a deliberation part between the proposition and the decision.

The clerks of the States General used formulas to signal important elements of resolutions: the start of a proposition, the type of proposition, the start of a decision, the type of decision, a request for a committee to investigate a matter further, or a reference to a previous resolution. Our aim is to algorithmically extract and classify these formulas, to provide additional metadata per resolution. This process is complicated because formulas contain lots of variation, caused by, amongst others:

- Common linguistic variations due to grammatical changes (single, plural, present tense, past tense etc), orthographic variability and OCR/HTR mistakes,
- Variations in writing styles between clerks resulting in slightly different wording,
- Changes in the entities (often dates, names of persons and locations) contained within a formula,

- Changes due to qualifications of the formula, sometimes also extending the formula.

In Table 1 we have included a non-exhaustive example of the formula variation of one of the formulas that indicates a decision: "is goetgevonden en verstaen dat" (EN: 'is approved and understood that'). Even if the example is one of the more standardised formulas in our corpus, it contains instances of all the reasons for formula variation summed up above. In most cases the formula will have been preceded by "waer op gedelibereert zijnde" (EN: 'on which having deliberated') and they will of course be followed with an indication of what is approved.

Formula
is goetgevonden ende verstaen dat
zynde is goedtgevonden ende verstaan
goetgevonden ende verstaen dat copie
waer op gedelibereert zijnde is goetgevonden ende verstaen
ende verstaen dat een pasport
goetgevonden ende verstaen dat de
goetgevonden ende verstaen mits desen
is goedtgevonden ende verstaan dat aan
voorgaende deliberatie goedtgevonden ende verstaen dat
naer voorgaende deliberatie goetgevonden ende
is goedtgevonden ende verstaan mits
ende verstaen dat aen de
ende verstaen dat de heeren
goedt gevonden ende verstaan dat
is goedt gevonden ende verstaan
te werden en is dien onvermindert goedtgevonden ende verstaan dat
en is dienonvermindert goetgevonden ende verstaen dat
is goetgevonden en verstaen mits desen
ende verstaen dat de retroacta
goetgevonden ende verstaen dat gemelden
ende verstaen dat het voorschreve
voorgaende deliberatie goetgevonden ende verstaen dat pasporten in
voorgaande deliberatie goetgevonden ende verstaan mits desen
voorschreve waer op gedelibereert zynde is goetgevonden en verstaen
goedtgevonden ende verstaan mits dezen te
tot dordrecht is na voorgaende deliberatie goetgevonden en verstaen dat
ende verstaen dat het collegie ter admiraliteyt
goetgevon den ende verstaen dat
voorgaende deliberatie goetgevonden en verstaen dat pasporten ...

Table 1: Example of formula variation 'goetgevonden en verstaen dat'

The example shows that except for the most standard formulas, variation makes generalising formula form and detecting and grouping formulas very challenging. In many cases the only common elements are function words, but these are too common to use for finding formulas among a large corpus.

Class	# of formulas	# of occurrences
Total	100	2,995,794
Proposition	6	333,490
Decision	30	1,303,246
Resolution	9	208,943
Location	7	127,317
Variable	48	1,022,798

Table 2: Classes of formulas and their frequencies

2 Algorithmic Formula Detection

We are developing a computational method for identifying formulas in the corpus that is able to cope with variations. For our purposes the whole formula is of importance. Methods that have been developed in a computational phraseology (Pastor and Colson, 2020, Wahl and Gries, 2020, Wible et al., 2006) context are not suitable for our purposes, as they perform poorly on the many types of variation in the formulas of the resolutions.

Instead, like the Adjusted Frequency List algorithm by O'Donnell, we employ an algorithm that uses word n-gram frequencies, but includes steps to group variants and extend fixed-length n-grams to longer formulaic sequences using transitional probabilities (Koolen and Hoekstra, 2022).

We applied our algorithm on the printed resolutions, which covers the period 1705-1796, and consists of 286,340 resolutions.[1] This resulted in a list 7,055 formulas, many of which are variants or extensions of each other. Some of these formulas were known to us in advance and have already been used to identify the start of a proposition and thereby also the start of the resolution that it is part of. But the list also contains many formulas that we had not encountered before or had not considered as being formulaic, and some of these formulas express valuable information for providing digital access.

3 The location and function of formulas

To better understand the function of formulas, we focus on the top 100 most common formulas and analyse their location in resolutions and the context in which they appear.

We developed a classification scheme for determining the function of formulas and classified the 100 most common formulas, to understand how formulas were used to structure the work of the SG and the growing archive of previous decisions. We identified five different classes (see Table 2): formulas that signal aspects of the *proposition*, or of the *decision*, formulas that identify a reference to a (*previous*) *resolution* or to a (*geographic*) *location*, and *variable* formulas that can be used in multiple contexts. An example of a *variable* formula is "gedeputeerden van de provincie van" (EN: 'deputies of the province of'), which is followed by the name of one of the seven provinces of the Dutch Republic. Its function depends on where it occurs in the resolution, which is also variable. It can be near the start when the deputies submit a proposition, or anywhere in the decision paragraph, if the decision involves a particular province.

We used fuzzy search[2] to find all occurrences of these 100 formulas in the printed

[1] The handwritten resolutions are not available yet.

[2] Using our own Fuzzy-search module, see `https://pypi.org/project/fuzzy-search/`.

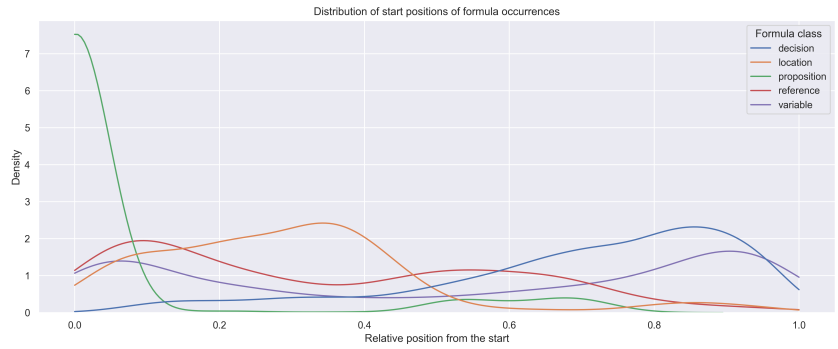

Figure 1: Distribution of the character position of the top 100 formula occurrences relative to the start of the resolutions they occur in. On the X-axis, 0.0 is the start of the resolution, 1.0 the end.

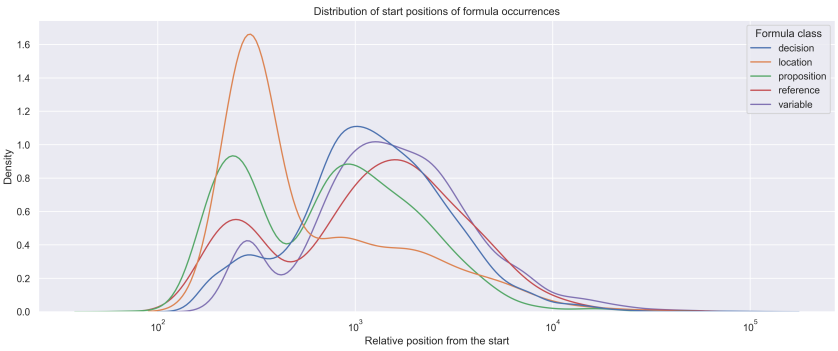

Figure 2: Distribution of the length of resolutions in which the top 100 formulas occur.

resolutions, resulting in almost 3 million matches. Their distribution across the classes is shown in Table 2. The frequencies of the individual formulas is highly skewed, with the top few formulas having hundreds of thousands of occurrences, but many having a few (tens of) thousands of occurrences.

Next, we link the formulas to their location in the resolution text, and find that some formulas have stable positions, occurring at the same part of a resolution across hundreds of thousands of resolutions, while others have a more variable location. The distribution of formulas, grouped by their class, is shown in Figure 1. Unsurprisingly, *proposition* formulas mostly occur at or near the start of resolutions, while *decision* formulas tend to occur in the second half of resolutions. Formulas relating the resolution to a geographic *location* tend to occur in between *proposition* formulas and the *decision* paragraph. In most cases, the *location* formula indicates where a proposition was sent from. These findings suggest that we can use location to semi-automatically classify the remaining 6,955 formulas.

Furthermore, we find that the function of formulas is related to the length of resolutions (see Figure 2). The *location* formulas tend to appear in short resolutions based on missives, where the location of the representative diplomat is always mentioned. Many of these missives contain information updates that require no decision, resulting in a short resolution.

We also find that sets of formulas tend to co-occur, signalling classes of relationships that can provide additional information about the content and context of individual resolutions as well as sequences of related resolutions. The formula "en van alles

alhier ter vergaderinge rapport te doen" (EN: 'and to report on everything here in the meeting') only occurs in combination with the formula "gestelt sal werden in handen van" (EN: 'will be put into the hands of'). The latter is a formula to signal that a decision has been postponed and that a committee is formed to further investigate a matter. The former signals that the committee is to report back their findings to the SG.

3.1 Formulas and Resolution Types

In earlier work, we identified 32 formulas that introduce a proposition, and thereby signal the start of a resolution Koolen et al. (2020). Using these opening formulas, we have made a provisional classification (Table 3) of proposition types for the printed part of the corpus. Each resolution has a single proposition type. The distribution of the top 100 formulas across the resolutions per proposition type is also shown in Table 3. Note that some of the formulas partially overlap with each other, so the total number of non-overlapping formulas per resolution is lower. Resolutions based on missives have fewer formulas (9.0 on average) than resolutions based on petitions (11.9) or reports (17.7). Reports represent resolutions that address a postponed decision from an earlier resolution.

Proposition Type	Number	Percentage	Formula occurrences	
			Total	Per resolution
Total	286340	(100%)	2,995,794	10.5
Missive	161,150	(56%)	1,448,207	9.0
Petition	76,454	(27%)	909,435	11.9
Report	13,083	(5%)	231,917	17.7
Unknown	10,173	(4%)	104,834	10.3
Memo	9,393	(3%)	106,895	11.4
Oral	8,409	(3%)	86,745	10.3
Resolution	2,544	(1%)	38,843	15.3
Conclusion	1,736	(1%)	8,796	5.1
Declaration	1,291	(0%)	11,731	9.1
Recommendation	1,061	(0%)	14,129	13.3
Bill	714	(0%)	6,723	9.4
Instruction	257	(0%)	678	2.6
Passport	75	(0%)	633	8.4

Table 3: The number of resolutions per proposition type, and the number of occurrences of the top 100 formulas.

4 Discussion

Formulas are used throughout the resolution corpus, more or less consistently, providing a useful starting point for extracting information about individual resolutions. But the large number of formulas and the amount of variation poses a number of challenges to derive meaningful dimensions for information access.

One challenge is how to make the entire formula detection and classification process manageable. The number of resolutions and formulas is big, and there is no clear

boundary between what is a formula and what is not, and between formulas that are useful for improving information access and those that are not. We currently take an iterative approach. We start with identifying the most frequent formulas, and structure the content of the resolutions (e.g. identify the boundaries between proposition and decision paragraphs). Then we repeat the process, using the more structured resolutions as input, to have more context to identify and categorise further formulas. With each successive step, we expect the formulas to become less frequent, more variable and less informative. Therefore, there are diminishing returns with further iterations. What is a good number of iterations needs to be determined experimentally.

Another open question is how we can establish a useful definition of and model for formulas that handles variation but gives a clear and unambiguous meaning.

Finally, we will investigate how we can best incorporate variable information such as entity names into formulas.

References

Ute K Boonen. De begin-en slotformules in utrechtse oorkonden uit de dertiende en veertiende eeuw: een vergelijking van middelnederlandse en latijnse formuleringen. *Neerlandistiek*, 2005, 2005.

Helmut De Boor. *Actum et datum: eine Untersuchung zur Formelsprache der deutschen Urkunden im 13. Jahrhundert*. Number 4. Verlag der Bayerischen Akademie der Wissenschaften, 1975.

Marijn Koolen and Rik Hoekstra. Detecting formulaic language use in historical administrative corpora. In Folgert Karsdorp and Kristoffer L. Nielbo, editors, *Proceedings of the Computational Humanities Research Conference 2022, CHR 2022, Antwerp, Belgium, December 12-14, 2022*, volume 3290 of *CEUR Workshop Proceedings*, pages 127–151. CEUR-WS.org, 2022. URL `http://ceur-ws.org/Vol-3290/long_paper5740.pdf`.

Marijn Koolen, Rik Hoekstra, Ida Nijenhuis, Ronald Sluijter, Esther van Gelder, Rutger van Koert, Gijsjan Brouwer, and Hennie Brugman. Modelling resolutions of the dutch states general for digital historical research. In *COLCO*, pages 37–50, 2020.

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

Joanna Kopaczyk. *The legal language of scottish burghs: standardization and lexical bundles (1380-1560)*. Oxford University Press, 2013.

Claire Lemercier and Francesca Trivellato. 1751 and thereabout: A quantitative and comparative approach to notarial records. *Social Science History*, 46(3):555–583, 2022.

André Evangelista Marques. Between the language of law and the language of justice: The use of formulas in portuguese dispute texts (tenth and eleventh centuries). In *Law and Language in the Middle Ages*, pages 128–164. Brill, 2018.

Matthew Brook O'Donnell. The adjusted frequency list: A method to produce cluster-sensitive frequency lists. *ICAME Journal*, 35:135–169, 2011.

Gloria Corpas Pastor and Jean-Pierre Colson. *Computational Phraseology*, volume 24. John Benjamins Publishing Company, 2020.

Theo Thomassen. *Onderzoeksgids: Instrumenten van de macht: de Staten-Generaal en hun archieven 1576-1796* (*Band 1*). Sidestone Press, 2019. ISBN 9789088908798.

Alexander Wahl and Stefan Th Gries. Computational extraction of formulaic sequences from corpora. *Computational Phraseology*, 24:83, 2020.

David Wible, Chin-Hwa Kuo, Meng-Chang Chen, Nai-Lung Tsao, and Tsung-Fu Hung. A computational approach to the discovery and representation of lexical chunks. In *TALN*, pages 868–875, 2006.

Amalia Zomeño. Notaries and their formulas: The legacies from the university library of granada. In *From al-Andalus to Khurasan*, pages 59–77. Brill, 2007.