A Corpus of Early Modern Decision-Making - the Resolutions of the States General of the Dutch Republic

Marijn Koolen and Rik Hoekstra
DHLab - KNAW Humanities Cluster
Huygens Institute
Amsterdam, Netherlands
marijn.koolen,rik.hoekstra@di.huc.knaw.nl

Abstract

This paper presents a corpus of early modern Dutch resolutions made in the daily meetings of the States General, the central governing body of the Dutch Republic, over a period of 220 years, from 1576 to 1796. This corpus has been digitised from over half a million scans of mostly handwritten text, segmented into individual resolutions (decisions) and enriched with named entities and metadata extracted from the text of the resolutions. We developed a pipeline for automatic text recognition for historic Dutch, and a document segmentation approach that combines ML classifiers trained on annotated data with rule-based fuzzy matching of the highly formulaic language of the resolutions. The decisions that the States General made were often based on propositions (requests or proposals) submitted in writing, by other governing bodies and by citizens of the republic. The resolutions contain information about these submitted propositions, including the persons and organisations who submitted them. The second part of this paper includes an analysis of the information about these proposition documents that can be extracted from the resolutions, and the potential to link the resolutions to their corresponding propositions using named entities and extracted metadata. This will allow historians and genealogists to study not only the decision making of the States General in the early modern period, but also the concerns put forward by both high-ranking officials and regular citizens of the Republic.

1 Introduction

We present the dataset of resolutions of the States General of the Dutch Republic (from 1576 until 1796), a corpus of 692,712 resolutions (decisions) and 130 million words of early modern Dutch text. In addition, we release a dataset of almost 8 million associated entity mentions, and at least partial disambiguation and linking to entities for 5 million of these mentions.[1] The resolutions are the decisions taken by central governing body of the Dutch Republic, which were written down as minutes during daily meetings for a period of 220 years, and extended and recorded in resolution books.

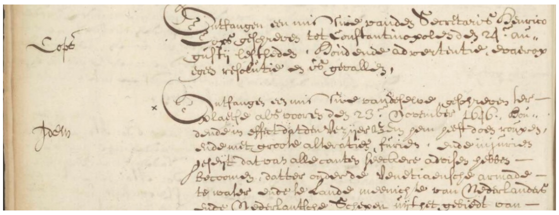
This corpus is of great relevance to researchers for multiple reasons. First, for political historians, the States General (SG) of the Dutch Republic is an important example of early modern republic decision making, and the long period allows researchers to trace the different steps in decision making processes as well as analyse the interaction of the SG with foreign powers, regional organisations and individual citizens. Second, for linguistics, the consistent recording of decisions during daily meetings for 220 years, by a relatively small group of clerks, represents a great resource for studying changes in spelling, word choice and syntax in a specific domain. Third, the resolutions were made in response to requests or proposals submitted to the SG, mostly in written form, and these documents have been archived and recently digitised as well. This offers an opportunity to link the resolutions to the archive of correspondence send to the SG and improve accessibility to those documents.

In this paper, we describe how the corpus was constructed, analyse which organisations and individuals send requests or proposals to the SG to understand how the SG interacted with different groups in society, and with that, explore the potential for linking the resolutions to the archive of incoming correspondence.

Each resolution consist of a proposition (a request or proposal submitted to the SG, mostly in written form) and the decision on that proposition. An example handwritten resolution, made on 28 January 1647, is shown in Figure 1, in which a secretary, Henrico Cops, has send two missives[2],

[1] The data is also available via our online search application Goetgevonden, https://app.goetgevonden.nl

[2] A missive is letter sent between two officials.



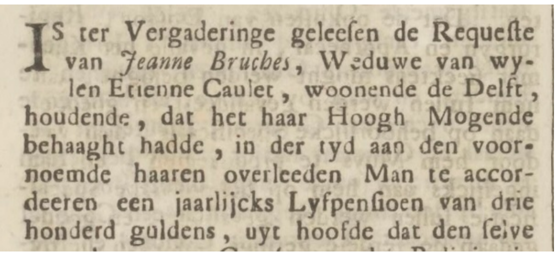
Cops

Ontfangen een missive vanden Secretaris Henrico
Cops geschreven tot Constantinopolen den 24e. au=
gustij lestleden, houdende advertentie, waerop
egeen resolutie en is gevallen,

Idem

× Ontfangen een missive van deselve, geschreven ter—
plaetse als vooren den 23e. November 1646., hou=
dende in effect dat den Vezijer Bem hem heeft doen roupen,
ende met groote alteratien, furien, ende injurien
geseijt dat van alle canten seeckere advisen hebben —
becoomen, datter onder de Venetiaensche armade —
te water ende te lande menichte van Nederlanders
ende Nederlantsche Schepen uijt het gebiedt van —

Figure 1: Handwritten resolutions of 1647-01-28. Source: https://www.nationaalarchief.nl/onderzoeken/archief/1.01.02/invnr/3253/file/NL-HaNA_1.01.02_3253_0075



IS ter Vergaderinge geleeſen de Requeſte van *Jeanne Bruches*, Weduwe van wylen Etienne Caulet, woonende de Delft, houdende, dat het haar Hoogh Mogende behaaght hadde, in der tyd aan den voornoemde haaren overleeden Man te accordeeren een jaarlijcks Lyfpenſioen van drie honderd guldens, uyt hoofde dat den ſelve

Figure 2: First part of a printed resolution of 1756-09-01. Source: https://www.nationaalarchief.nl/onderzoeken/archief/1.01.02/invnr/3811/file/NL-HaNA_1.01.02_3811_0303

in August and November 1646 from Constantinople (current-day Istanbul). The transcription of the proposition paragraphs read:

> Ontfangen een missive vanden Secretaris Henrico Cops geschreven tot Constantinopolen den 24e. augustij lestleden, houdende advertentie, waerop egeen resolutie en is gevallen.
>
> (EN: *Received a missive of the Secretary, Henrico Cops, written in Constantinople, the 24th of last August, containing intelligence, on which no resolution was made*)
>
> Ontfangen een missive van deselve, geschreven ter plaetse als vooren den 23e. November 1646 houdende in effict dat den Vezijer Bem hem heeft doen roupen, ende met groote alteratien furien, ende injurien geseijt dat van alle canten seeckere advisen hebben becoomen, ...
>
> (EN: *Received a missive of the same, written on location as previous on the 23rd of November 1646, stating in effect that the Vizier Bem had summoned him and, with great agitation, fury, and insults, declared that they had received certain reports from all sides, ...*

No detail is provided about the content of the first missive and no decision is taken, but in the second missive, Cops asks for an 'ad omnes populos' (a passport) for his housekeeper, which is granted in the decision part (not shown).

An example of a printed resolution, taken on 1 September 1756, is shown in Figure 2, in which the proposition is a 'Requeste' (petition) by Jeanne Bruches, widow of Etienne Caulet, living in Delft, requesting a annual pension of 300 guilders.

The resolutions have a very regular structure and contain many formulaic phrases (Thomassen, 2019b; Koolen et al., 2023a), which allow us to algorithmically extract metadata about each resolution. This includes the date of the resolution and how the proposition was submitted, e.g. orally during the meeting, or in writing. For written proposals, the resolutions mention the type of document—e.g. a missive or a petition as in the examples above, but there are various other types—which is strongly related with the status of the proposer. Missives are always associated with formally appointed representatives of the SG, while petitions could be submitted by anyone.

We also tagged the corpus of resolutions with seven types of entities, including four common types—persons, locations, organisations and dates— and three domain- and corpus-specific types: committees, references to earlier resolutions and person attributions. Committees were small groups of persons selected from the members of the States General and tasked to investigate a matter and report back before a final decision is taken. References to earlier resolutions represent a link between two resolutions. Together, the committees and references allow one to trace the chain of decision making around specific matters. Finally, person attributions are part of person entities that have been tagged separately to separate person proper names from other identifying attributions like professions (e.g. carpenter, ship captain, ambassador or lawyer), titles (duke, earl or queen) and legal status (e.g. daughter, son, widow, minor, orphan or heir)

Outside of our project, the incoming correspondence of written propositions has been digitised as well,[3] prompting us to analyse whether we can use the patterns in the resolutions about the proposer and proposing document that can help in linking the resolutions to their corresponding proposition

[3] For the incoming correspondence, this is currently limited to scanning the physical documents, and generating transcriptions using a generic ATR model trained on early modern Dutch texts. No document segmentation or metadata extraction has been done.

document.

We address the following research questions:

- How can we combine machine learning and domain knowledge about formulaic language use for document segmentation and metadata extraction?
- Can we identify patterns of entities and formulas in the resolutions that are related to propositions?
- Can the categorisation of person attributes give insight in what groups of people engage with the SG over time?

In this paper we make the following contributions. First, we publish the resolutions as long serial corpus with rich metadata as an Open Access dataset. Second, we analyse proposition patterns to show that most propositions are submitted by single persons or organisations, mostly from the domains of politics and administration, but that there are tens of thousands of propositions made by regular citizens of the Republic. And third, we show that, because of the highly standardised format of the resolutions, we can extract metadata from the majority of resolutions about the proposition, which can help us link them to the submitted proposition documents.

2 Related Work

The resolutions of the SG only reflect the decision making process. Thomassen calls the SG a decision making machine. Final decisions were usually assumed to be taken unanimously, and the deliberations were not recorded (Thomassen, 2019a, p.101,196). This contrasts with records of parliamentary debates, which contain not just the final decisions, but also the deliberations, that is, the exchange of political arguments. Well-known examples of modern parliamentary debates are the Parlamint corpora[4] (Erjavec et al., 2023), but there are also more historical corpora of debates (Hyvönen et al., 2025; Puren et al., 2025).

There is a long tradition in publishing parliamentary deliberations, decisions and associated papers, that started in the 19th century in book form. More recently, many of these books have been digitised and sometimes extended for better access (see Hoekstra et al. (2025) for an overview of earlier editions of the resolutions). In a continuation of previous book publications of the *Reichstagsakten*, Bleier et al. (2023) published a digital edition of the Regensburger Reichstag, manually transcribing texts of documents and encoding elements of communicative acts (senders, receivers, decision makers, decisions, etc.) using CIDOC CRM.

Several digitisation projects of early modern parliamentary documents are presented in (Zeilinger et al., 2025), e.g. the digitisation of documents that recorded the activities the Polish Seym or Diet in the 16th century (Fokt and Mikuła, 2025), and of early modern English parliamentary acts, bills and other documents (Seaward and Matwin, 2009). One of the most extensive digital publications are the records of the parliaments of Scotland to 1707,[5] “a fully searchable database containing the proceedings of the Scottish parliament from the first surviving act of 1235 to the union of 1707." It is compiled from manuscript sources, earlier editions and additional archival materials. There are undoubtedly more examples, but as far as we know there are no up-to-date overviews of such collections and initiatives.

3 The Corpus of Resolutions

We published the corpus of Resolutions of the States General of the Dutch Republic as an Open Access resource on Zenodo (Koolen et al., 2025a).[6] This long, serial publication is relevant for research on, amongst other, linguistics, political history and political science.

Each resolution contains at least a *decision* paragraph. From 1637, all resolutions consisted of three parts: (1) a paragraph describing a *proposition* submitted in a written document—e.g. a missive, petition, letter, report, bill or memo—and some details of what was proposed or requested, (2) a decision paragraph detailing what action is to be taken—which can include postponing a decision and asking for advice first, or to not take any action—and (3), a decision making formula that connects the two. This formula is a more-or-less fixed phrase, *“Waerop gedelibereert synde, is goetgevonden ende verstaen dat”* (EN: *Upon deliberation, it has been accepted and understood that*). However, even once the formula was more-or-less fixed, still variations occurred, either intentionally

[4] See also `https://www.clarin.eu/parlamint`

[5] `https://www.rps.ac.uk/`

[6] Currently only in TSV format with plain text paragraphs and metadata. In the next months we will add richer formats with additional metadata.

or unintentionally (omitting parts, changing word order or using different spelling).

Each resolution has metadata including the date on which the decision was taken, the type of proposition, and whether it was an *ordinary* or a *secret* resolution. Many decisions relating to e.g. matters of war or political negotiation were considered sensitive and recorded in separate secret resolution books, which were only accessible a small group of actors.

3.1 Corpus Construction

The construction of the corpus followed a number of automated and manual steps.

Text recognition The first step was making transcriptions of the text in the 278,872 scans from 657 books. For the ordinary resolutions of 1703-1796, we used the available printed volumes. All earlier resolutions and the secret resolutions of 1703-1796 are only available in handwritten versions. The transcriptions were produced by Loghi,[7] an end-to-end layout analysis and Automatic Text Recognition (ATR) pipeline that we developed in the context of this project and other projects (van Koert et al., 2024), which handles both handwritten and printed text. The ATR pipeline consist of multiple steps. LayPa is used for layout analysis for baselines detection and text line segmentation (Klut et al., 2023). Next, the text is recognised using convolutional and recurrent layers, and the output layer is passed through Connectionist Temporal Classification (Graves et al., 2006) to find the most likely sequence of characters (van Koert et al., 2024). ATR on the printed volumes required ground truth transcriptions for only a small set of 107 scans to reach a Character Error Rate (CER) of 1%. For the handwritten texts, the ground truth consists of 515 scans with which we reached a CER of 3%. Both ground truth datasets are published on Zenodo van Koert, 2023; Sluijter et al., 2023.

Document segmentation The next steps was document segmentation, for which we detected paragraph boundaries (taking into account that paragraphs can cross page boundaries), and then use formulaic phrases to identify whether a text line is the start of a *meeting* or a *resolution* or not. Pages consist of different elements. The resolutions taken on the same day are preceded by the date of the meeting and a list of attending SG members. Next to most handwritten resolutions there are one or more marginalia (terms describing the content of the resolution) written by a clerk, which was later copied in indexes together with a page reference. Finally, many pages contain headers and footers such as page numbers, dates and catch words.[8]

	Number		
Element	Total	Printed	Handwr.
Sessions	108,802	32,675	76,127
Resolutions	692,156	304,710	387,446
Entities	8,032,123	4,523,248	3,508,875
Person name	1,895,298	1,089,223	806,075
Attribution	2,194,178	1,348,829	845,349
Organisation	601,648	330,955	270,693
Committee	141,396	75,964	65,432
Location	2,167,993	1,212,529	955,464
Date	844,337	379,573	464,764
Resol. ref.	187,273	86,175	101,098

Table 1: Descriptive statistics of the elements in constructing the corpus

	Handwritten			Printed		
Element	#	P	R	#	P	R
Session start	56	1.0	0.86	76	1.00	0.99
Session date	56	0.96	0.82	76	0.99	0.97
Res. start	313	0.94	0.87	689	0.97	0.95

Table 2: Evaluation of identifying the start of a session, the date of a session and the start of a resolutions in terms of precision (P) and recall (R) for printed and handwritten texts

The segmentation process thus contains two steps: 1) segmenting the text of pages into text per meeting, and 2) segmenting the text per meeting into text per resolution. Next to meeting segmentation, the date of each meeting needs to be assigned. Below we describe the pipeline and the evaluation for these three tasks.

For the printed resolutions, distinguishing between text lines that are part of the resolutions and text lines that are headers (page numbers or dates) or footers (catch words) is done in the ATR pipeline (text lines are grouped into regions that are classified as resolution, header or footer). In Koolen et al. (2023b) we describe our approach and evaluation for identifying the start of a meeting and the exact

[7] `https://github.com/knaw-huc/loghi-htr`

[8] Catch words repeat the first word of the next page to help check that a multi-page sheet is folded correctly, see `https://en.wikipedia.org/wiki/Catchword`

date (see also Table 2). Because we introduced improvements to our formula detection model based on the evaluation reported in that paper, we decided to create a new ground truth set of 200 randomly sampled printed pages and identify all starts of meetings (76 in total) and of resolutions (689 in total) and the dates of those 76 meetings. The evaluation results of our approach on printed pages is shown Table 2. Both precision and recall are close to 1.0, indicating that few mistakes are made. This is due to the extremely consistent manner in which the resolutions were printed. The few mistakes are mostly due to bad transcriptions because of damaged pages and a few exceptions where a resolution does not start with a fixed formula.

For the handwritten resolutions, which is the majotity of the material, this process is less straightforward, as the layout is less consistent and elements are not always clearly visually separated. This is partly due to resolutions being written by many different scribes, who occasionally wrote all elements very compactly to use paper efficiently, but which makes segmentation more challenging.

The ATR pipeline classifies lines as either *resolution*, *meeting date*, *attendance list*, *marginalium* or *page number* but makes many mistakes. To alleviate this problem, we used Flair (Akbik et al., 2019) to train a Bi-LSTM sequence tagger that incorporates both textual features of each text line and visual-spatial features (e.g. coordinates of the line on the scan) to classify the sequence of text lines on a page.[9] We manually tagged 13,525 text lines from 332 randomly selected pages of handwritten text.[10] For the textual features we experimented with contextual character and word embeddings (separately and in combination). We used Flair Contextual Embeddings (Akbik et al., 2018) to train forward and backward character embeddings on the corpus of resolutions. For word-level embeddings we used GysBERT (Manjavacas and Fonteyn, 2022), which is a model trained on historic Dutch. The best model uses character embeddings, GysBERT and visual features, and is effective in distinguishing the main classes of text lines (paragraphs, meeting dates, lists of attendants and marginalia) with an overall accuracy of 0.92, but less effective at distinguishing between the start, middle or end of a paragraph (accuracy 0.82) and despite the highly formulaic language of the resolutions, much less at identifying lines that are the start of a resolution (accuracy 0.67). Finally, we use lists of known formulaic phrases that signal either the start of a proposition (and therefore the start of a resolution) or the date of a meeting. We use FUZZY-SEARCH,[11] a fuzzy search module we developed to search texts for known phrases occurring with variations in spelling or with ATR errors. The fuzzy searcher uses a dictionary of phrases as input and searches paragraphs for any possible occurrence of a known phrase using a similarity threshold based on Levenshtein distance and the length of the phrase.

For the start of a proposition we use a list of 188 formulas (see next paragraph for more details). Some of these formulas were known in advance, others were algorithmically detected (Koolen and Hoekstra, 2022). For the meeting dates we use a list of between 20 and 60 dates—in several different date formats—around the date of the previously found meeting date, as the meeting dates are mostly chronologically recorded and we therefore expect the next date to be shortly after or before the previous date. This is very sensitive to the amount of variation with which these meetings were recorded. In the early years, there is more variation, in terms of: formatting of dates (with or without the names of weekdays in Latin, with or without an infix like 'den' (EN: the) between weekday and day of the month, using Roman or Arabic numerals or Latin dative for the day of the month, the names of months in Latin or Dutch, with or without abbreviations, with or without the year, etc.), in terms of the chronology (sometimes additional resolutions of an earlier date are recorded after the resolutions of a later meeting, so there are two sections with resolutions from the same date, with resolutions of a later date in between) and in terms of the gap between meetings (mostly one or a few days, but sometimes multiple weeks). In the last roughly 120 years, the meetings and their recordings were highly consistent, as far as we have been able to establish being completely chronological and rarely skipping more than three days, usually around known and predictable, and thus computable, holy days like Easter, Pentecost and Christmas.

[9] The codebase of the post-ATR processing pipeline is available on Github (see https://github.com/HuygensING/republic-project) and published on Zenodo (Koolen et al., 2025b).

[10] These 332 pages were sampled independently of, and two years before, we sampled the 200 pages for evaluation.

[11] https://github.com/marijnkoolen/fuzzy-search

In this segmentation step, we use fuzzy search to identify lines in the transcription that are the start of a resolution or of a meeting. When the fuzzy searcher finds a proposition formula, the text line is classified as both the start of a paragraph and of a resolution. When it finds a meeting date, the line is classified as the start of a meeting. When it finds no formula, we use the classifications provided by the sequence tagger, as it has higher accuracy than the ATR pipeline.

To evaluate how well we can find the start of a meeting, the correct date of that meeting and the start of a resolution, we manually tagged 313 resolution starts and 56 session starts in a random sample of 200 pages of the handwritten resolutions. For identifying the start of a meeting, our approach attains a precision of 1.0 and recall of 0.86 (48 out of 56 meeting starts). For identifying the correct date, we reach an precision of 0.96 when a meeting start is found (46 out of 48 starts) and a recall of 0.82 for all meeting starts (46 out of 54 starts). Because the resolutions are ordered mostly chronologically, a consequence of not recognising the start of a meeting is that the resolutions of that meeting are assigned the same date as the previous meeting, which is usually the day before, or in the case of Mondays, two days before, since there normally was no meeting on Sundays.

For identifying the start of a resolution, precision is 0.94 and recall is 0.87. Although not as accurate as for the printed resolutions, the overwhelming majority of resolutions are correctly segmented and assigned the correct date.

After the segmentation steps, we end up with a corpus of 692,156 resolutions, each assigned to one of 60,046 days. Resolutions vary strongly in length, in terms of number of words per resolution, from a single word (due to incorrect segmentation) to 56,353 words. The distribution is skewed, but the median resolution has 117 words and the interquartile range is (25-75%) between 47 and 242 words and 90% of all resolutions have between 23 and 684 words.

Proposition type extraction The 188 known formulaic phrases vary in length between 3 and 15 words, and many of them contain a term to describe what type of document was submitted to the SG, as a proposition that gave rise to the resolution.

For example, the phrases *“Ontfangen een Missive van ...”* (EN: *Received a Missive of ...*) and *“Is ter Vergaderingen gelesen de Requeste van ...”* (EN: *Has been read during the Meeting the Request of ...*), are two fixed formulaic openings introducing a missive and petition respectively on which resolutions were made, including the resolutions in Figure 1 and 2. These formulas occur in the corpus, with some spelling variation, tens of thousands or even hundreds of thousands of times (see Table 3).

Formula	**Count**	**Frac.**
Ontfangen een *Missive* van	284,703	0.41
Is ter Vergaderinge gelesen de *Requeste* van	95,910	0.14
Op de *Requeste* van	46,780	0.07
Is gehoort het *Rapport* van	26,083	0.04
Is ter Vergaderinge gelesen de *Memorie* van	19,019	0.03
Is goetgevonden	17,063	0.02
Synde ter Vergaderinge gelesen de *Requeste* van	12,441	0.02
heeft ter Vergaderinge voorgedraagen	9,707	0.01
Ontfangen eenen *brieff*	8,276	0.01
Other (179 formulas)	111,581	0.16
No formula	60,593	0.09

Table 3: Frequency of the most common formulaic phrases for starting a resolution. The proposition document types are highlighted in italics.

These document types carry information about what kind of proposition was submitted (Riemsdijk, 1885; Thomassen, 2019a,b). Proposition types have clear definitions and are consistently used in the resolutions to describe the documents sent to and by the SG (Thomassen, 2019b, pp.807–820). The distribution of proposition types is shown in Table 4. A *missive* is a letter from one authority to another, written in first person. They were mostly submitted by diplomats and ambassadors of other governing bodies, both in the Republic and abroad. A *petition* is a request that could be submitted either by government authorities or by citizens of the republic. *Reports* were submitted by committees of the SG, consisting of SG members who were tasked to investigate matters if the SG deemed more information was needed before making a final decision. The term *letter* was used until 1637 for a broad category of documents with an explicit sender and receiver. Most of these letters would after 1637 be referred to as missives. A *remonstrance* is a petition from a lower governmental layer to a higher one. There are 15 other proposition types in the corpus. The *orally* introduced propositions (3%) have no archived document associated with them. For 14% of the resolutions the proposition type cannot be derived from the formula, so we do not know if there is an archived proposition document. That means that for at least 83% of the resolutions, the proposition document should be in the archive of received documents, and the combination of proposition type and named entities could be used to link to their corresponding resolutions.

Proposition type	freq.	%
Missive	293,823	0.42
Petition	177,962	0.26
Unknown	99,046	0.14
Report	32,493	0.05
Memo	24,243	0.04
Oral	19,655	0.03
Letter	11,824	0.02
Resolution	11,784	0.02
Remonstrance	7,044	0.01
Other	14,282	0.02
Total	692,156	1.00

Table 4: Distribution of proposition types associated with at least 1% of all resolutions.

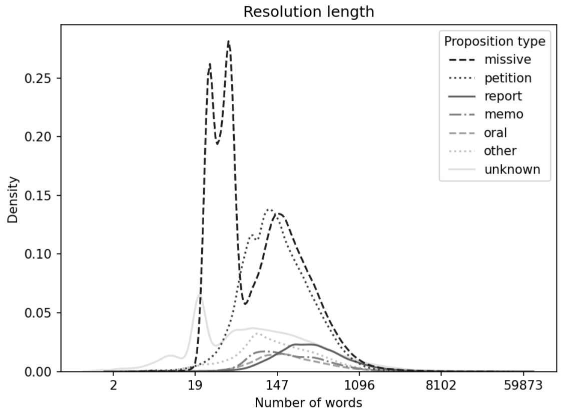


Figure 3: Distribution of resolution length in number of words for resolutions based on different types of propositions.

With the proposition types identified, we can see that resolutions based on different proposition types have different length characteristics (Figure 3). Resolutions based on *missives* are relatively short, with a median of 70 words, and the distribution is bi-modal. The first peak corresponds to resolutions where the missive did not lead to any decision, ending with the formula *"Waarop geen resolutie is gevallen"*, (EN: *on which no resolution was made*.) Resolutions based on reports are the longest on average (median is 317 words and 90% are longer than 120 words).

Entity recognition and disambiguation For training a NER tagger, we used Flair (Akbik et al., 2019) and multiple types of embeddings, including Flair contextual character embeddings (Akbik et al., 2018) and FastText embeddings (Bojanowski et al., 2017) that were trained on the corpus of resolutions, and GysBERT (Manjavacas and Fonteyn, 2022). Details of the training and evaluation of these NER models on the resolutions is described in Koolen et al. (2024, 2025c).

The entity data is published on Zenodo (Dijkstra et al., 2025) and contains mentions of seven types of entities, including the common types *person*, *organisation*, *location* and *date*, and three corpus-specific types *committee*, *resolution reference* and *person attribution*. Committees were part of the SG and consisted of members of the meetings who were tasked with investigating matters of a proposition before a final decision was taken. In such resolutions, the decision paragraph states that a given committee will investigate and report back. Example committees are *foreign affairs*, *maritime affairs* and *military affairs*. When a committee reported back to the SG (days, weeks or even months later), this was recorded as another resolution, with an explicit reference to the earlier resolution. This is tagged as a *resolution reference* by the NER tagger. Finally, a *person attribution* is any part of a person entity mention that is not a proper name, but describes their profession, a legal status, family relationship or title.

There are many nested entities in the corpus, which is common in historical corpora of administrative texts (Prada Ziegler, 2024; Aguilar et al., 2016). Person entities often contain smaller entities such as person attributions[12], organisations and locations. Committees often contain the person name of the chair of the committee (91% of committee entities have a person name) or a location (35%). The latter is the case for committees that deal with the affairs related to specific political regions.

The resolution in Figure 2 starts with *"Is ter Vergaderinge geleesen de Requeste van Jeanne Bruches, Weduwe van wylen Etienne Caulet, woonende de Delft, houdende, ..."* (EN: Has been read during the Meeting the Request of Jeanne Bruches, Widow of deceased Etienne Caulet, living in Delft). Here, *Jeanne Bruches* is the proper name, and the attribution is *Weduwe van wylen Etienne Caulet,*

[12] Attributions are strictly speaking not named entities when they are part of a person name, but they are tagged as such to allow categorisation.

woonende de Delft and together they are the person reference.

Sometimes, individual persons or groups are only referred to by an attribution, such as in the following formulaic start of resolution demonstrates : *"OP de Requeste van de gesamentlijcke Straatmaakers van het Hof"* (EN: On the petition of the combined bricklayers (lit. 'road pavers') of the court). There are no person names, nor are the bricklayers together an organisation. For such references, the tag *person attribution* is also used.

The process of resolving entity mentions to entities was done semi-automatically. In resolving entity mentions to entities, terms for locations, organisations and attributions such as professions, titles and status have been categorised (Koolen et al., 2024, 2025c). In the examples above, the person attribution term "Weduwe" (*widow*) has been categorised as *family relationship* and *Status and relationships*, and "Straatmaakers" (*bricklayers*) as *Profession* and *Trades, Crafts & Arts*.

4 Proposition Analysis

This section provides an analysis of the proposers who submitted documents to the SG. We first describe the number and entity types of proposers and how they relate to types of proposition documents. Second, we look at the syntactic structure of the formulaic phrases and how they can be combined with named entities to identify the proposers. Finally, we consider the categories assigned to entities to shed light on what kinds of persons and organisations submitted different types of proposition documents.

Number of Proposers The distribution of the number of persons or organisations involved in submitting a proposition is shown Table 5. The vast majority of proposition have only a single proposer. In the handwritten ordinary resolutions, there are relatively many propositions with multiple proposers, or with zero proposers, when compared to printed resolutions. Resolutions with zero proposers are cases where there is no explicit proposer information. In the majority of cases, there is a description of the proposition, but no proposer. In 17 out of 56 cases, there is no proposition, only a decision paragraph. In three cases, the proposition paragraph states that the proposer is the same as in the previous resolution.

# Proposers	Handwritten Ordinary		Printed Ordinary	
	#	%	#	%
0	50	0.16	6	0.01
1	226	0.72	639	0.93
2	32	0.10	31	0.04
3	2	0.01	9	0.01
4	2	0.01	2	0.00
5	0	0.00	1	0.00
6	1	0.00	0	0.00
8	0	0.00	1	0.00
Total	313	1.00	689	1.00

Table 5: The distribution of the number of persons or organisations who together submit a proposition. The first row, zero submitters, represents resolutions where no proposition is mentioned.

	Proposer type					
Proposition	Per	Grp	Com	Org	Multi	N/A
Missive	363	6	0	75	17	1
Memo	19	0	0	2	3	2
No type	13	2	7	3	2	6
Report	1	0	38	0	4	0
Petition	154	20	1	11	33	0
Resolution	1	0	0	20	0	3
Oral	11	0	0	14	1	3
Other	15	7	0	0	3	3
Total	577	35	46	125	63	18

Table 6: Distribution of proposer types per proposition type.

There is a temporal shift in the occurrence of resolutions with no clear proposer. Most of these occur before 1628, and then gradually decrease until around 1650, after which they are almost completely absent. From 1628, the SG issued an instruction that from then on, all propositions had to be submitted in writing, which slowly took effect (Thomassen, 2019a, p.162). Until 1650, there were still some ad hoc issues that arose during the meeting, on which the SG took a decision without a clear proposition. From around 1650, the meeting was completely formalised and virtually all issues were prepared in advance (Thomassen, 2019a, pp.122-123). This shift largely explains the differences we observe between handwritten and printed resolutions in Table 5. For the rest of the analysis we leave out this distinction and report on all resolutions combined.

The relationship between the type of proposer and the type of proposition is shown in Table 6. Because we only have proposition document type information from the document segmentation step, and not for all resolutions in the ground truth dataset, we use the 864 correctly recognised resolutions for the analysis. This skews the results to formulaic resolution starts we capture well, but in a manual sample of the unrecognised resolutions, we observed the same patterns.

The *Multi* column refers to all resolutions with more than one proposer. Most proposition types were submitted by a single person, and of these, the most common types were missives (363 out of 577, or 63%) and petitions (154 or 27%). When a group of persons submitted a proposition, it was most likely in the form of a petitions. These were often groups of merchants or ship owners asking for a passport to trade or sail abroad, or for the SG to weigh down on some conflict. Because committees were tasked to investigate a matter arising from a proposition and to report back, they were mainly the proposer or submitter of reports (38 out of 46, or 83%). Occasionally, they are the proposer of resolution where there is no clear proposition type (most likely, these should be classified as Oral). Organisations submitted missives (75 out of 125 or 60%) but also resolutions from other governing bodies (20 or 16%, most often resolutions taken by the States of the individual provinces).

Proposition formulas and entities Next, we look at the formulas and the recognised entities. Of the nine most common formulaic proposition phrases in Table 3, seven are syntactic constructions that introduce a proposition document as the subject of the sentence, to be followed by a direct object that is the proposer or submitter of that document. The same applies to 129 other proposition formulas. There are another 10 formulas where the proposer immediately precedes the formula. In other words, for 146 out of 188 formulas (78%), we can extract the immediately preceding or following entity mentions to extract information about the proposer. This information can help us link the resolution to the archived proposition document. In the entity resolution step, most organisations, committees and person attributions have been categorised, and we can use these categories to get an insight in what kinds of persons or organisations submitted what kinds of documents.

The distribution of proposer entity types per formula is shown in Table 7 for the most common formulas over the entire corpus of resolutions. This largely shows the same patterns as Table 6, which suggests that proposers are almost always recognised as entities and mostly categorised as the correct entity type.

This also means that for a large subset of the resolutions, we have at least one name of a person, committee or organisation to identify and link to the corresponding proposition document from the archive of incoming correspondence. Together with the date of the resolution and the short summary it provides of what was proposed, we can narrow down the possible candidate proposition documents in the correspondence archive, using these constraints.

For many of the submitted missives, we can go even further. When a missive was submitted, the formulaic phrase in the resolution typically contained the name of the proposer, the location and the date of sending it. The first resolution mentioned in Section 1 starts with the following formula: *<FORMULA>Received a missive of</FORMULA> the <ATT>Secretary</ATT>, <PER>Henrico Cops</PER>, written in <LOC>Constantinople</LOC>, the <DAT>24th of last August</DAT>, containing intelligence, on which no resolution was made*. As shown in Table 7, for 96% of the 284,703 resolutions that start with that formula, the NER tagger identified a person (name and or attribution) or organisation. Moreover, in 66% of these resolutions, a location is also identified, in 73% a date is identified, and in 65%, both a location and date are identified.[13] For these resolutions, we thus have multiple types of information for linking.

Finally, we look at the category labels of the recognised *person* entities and *person attributions* in combination with the proposition types, to get a better understanding of what kinds of persons engaged with the SG. Although it is to be expected that the majority of resolutions are based on propositions submitted by people involved in politics or administration and by nobility and rulers, it is valuable to know that citizens with a wide variety of professions and with different legal status were able to make their case with the SG.

Extraction Challenges and Evaluation In the analysis above, we used the syntactic structure of

[13] Dates without a location are found for missives send from The Hague, the same place as the SG meetings, when the resolution simply states that the missive was "geschreven alhier" (EN: written here)

		Entity type					No	
Formula	Pos.	PER	ATT	COM	ORG	LOC	Ent.	Total #
Ontfangen een *Missive* van	F	0.71	0.06	0.00	0.19	0.03	0.00	284,703
Is ter Vergaderinge gelesen de *Requeste* van	F	0.79	0.11	0.00	0.05	0.05	0.00	95,910
Op de *Requeste* van	F	0.87	0.10	0.00	0.01	0.01	0.00	46,780
Is gehoort het *Rapport* van	F	0.09	0.01	0.88	0.00	0.00	0.01	26,083
Is ter Vergaderinge gelesen de *Memorie* van	F	0.76	0.11	0.00	0.08	0.02	0.03	19,019
Synde ter Vergaderinge gelesen de *Requeste* van	F	0.76	0.17	0.00	0.04	0.03	0.00	12,441
heeft ter Vergaderinge *voorgedraagen*	P	0.46	0.07	0.02	0.36	0.04	0.05	9,707
Ontfangen eenen *brieff*	F	0.33	0.21	0.00	0.35	0.09	0.02	8,276

Table 7: The percentage of propositions where an entity directly follows (F) or precedes (P) a specific formula.

Attribution category	Missive	Petition	Unkn.	Report	Memo	Oral	Other	Total #
Politics & Administration	0.69	0.12	0.06	0.01	0.06	0.01	0.04	195,500
Nobility & Rulers	0.49	0.16	0.13	0.01	0.10	0.02	0.10	44,989
Army & Militias	0.35	0.48	0.08	0.01	0.03	0.01	0.05	38,180
Status & Relations	0.04	0.87	0.04	0.00	0.01	0.01	0.03	36,595
Trade, Crafts & Arts	0.01	0.89	0.03	0.00	0.01	0.00	0.06	26,398
Uncategorised	0.33	0.35	0.18	0.02	0.04	0.06	0.02	14,453
Agriculture, Shipping & Fishing	0.11	0.76	0.06	0.00	0.01	0.01	0.03	13,480
Legal	0.13	0.47	0.09	0.01	0.02	0.03	0.25	9,449
Religion	0.20	0.68	0.05	0.00	0.01	0.02	0.04	8,693
Finance	0.37	0.38	0.11	0.01	0.03	0.02	0.09	7,727
Services	0.02	0.86	0.07	0.00	0.01	0.01	0.03	4,198
Education & Research	0.05	0.65	0.08	0.00	0.01	0.03	0.18	2,397
Other	0.21	0.34	0.21	0.09	0.03	0.05	0.07	1,409
Geography	0.05	0.71	0.13	0.03	0.01	0.02	0.05	1,005

Table 8: The distribution of person attributions (in percentages) of proposers per proposition document type.

formulas to identify entities who are proposers submitting issues orally or via documents. Although the NER tagger almost always finds entities in the predicted positions, this does not mean that their entity type is correctly identified, that the entity boundaries are correctly detected, nor even that the identified entity is the actual proposer.

Moreover, as Table 5 shows, some 5-10% of resolutions have multiple proposers. In these resolutions, there are more complex patterns of proposition elements. When there are multiple proposers, especially proposers of different types (e.g. a person and an organisation or a person and a group), it is more challenging to correctly identify and extract the entities who are proposers. Even more challenging are the small number of cases where multiple documents are mentioned as the source of a proposition and decision. For instance, the following resolution mentions two missives as the source for the decision:

> ONtfangen een Missive van de Heeren haer Hoogh Mogende Gedeputeerden te Velde, als mede een Missive van den Heere Prince van Nassau, Erf-Stadthouder van Vrieslandt, en Stadthouder van Stadt en Lande, geschreven de eerste te Douay ende de andere te Leeuwaerden den vierden deser, houdende beyde antwoordt ...
>
> (EN: *Received a Missive of the gentlemen her High and Mighty Deputies in the Field, as well as a Missive of the Lord Prince of Nassau, Stadholder of Vrieslandt and Stadholder of Stadt en Lande, written the first in Douay and the other in Leeuwarden on the forth of this month, both holding response ...*)

We considered manually tagging only the proposers in our ground truth dataset, but, especially with multiple proposers and proposition documents, it remains difficult to determine the correctness of the entities and their assigned roles. Therefore, we decided it is better to postpone the evaluation and first develop a more complex conceptual model of proposition and create detailed annotation instructions and guidelines, so that we can tag both proposition documents, proposers, and entities who are explicitly mentioned as intermediaries (who pass on the document to the SG on behalf of the proposers), as well as the relationships between them—to make explicit, in the case of multiple propositions, which proposition is submitted by whom. Building a ground truth dataset with this more complex model will also allow us to train language models that can explicitly extract these roles and relationships.

5 Conclusions

In this paper we introduced the corpus of the Resolutions of the States General of the Dutch Republic (1576-1796) as an Open Access dataset of early modern political decision making. We described its construction in terms of text transcription, document segmentation, entity recognition and metadata extraction. Using a combination of machine learning and rule-based methods that employ domain- and corpus-specific knowledge and expertise, we were able to transform 657 books of handwritten and printed texts into a corpus of 692,712 individual resolutions (decisions) and automatically assign metadata about the proposition that forms the first part of most resolutions.

In addition, we conducted an analysis of the persons who submitted documents to the States General, which were in many cases the source of the resolutions. Our main reason for this investigation is that these corresponding documents have been digitised (but not yet segmented or enriched with metadata). Next steps include developing a more detailed ground truth set of proposition documents, proposers, intermediaries and relationships and train models to automatically extract the proposition information more explicitly.

Acknowledgments

This research is funded by the Dutch Research Council (NWO) through the NWO Groot project REPUBLIC (an acronym for REsolutions PUBlished In a Computational Environment) 2019-2024 (NWO grant number 175.217.024).

This work used the Dutch national e-infrastructure with the support of the SURF Cooperative using grant no. EINF-7089 and EINF-10206.

We would like to thank Esger Renkema, Nienke Groskamp, Frank Smit, Jirsi Reinders and Ger Dijkstra for curating and categorising the entity mentions. We thank the volunteers for their invaluable contributions to this project, including the creation and correction of tens of thousands of transcriptions of the resolutions, and annotation the entities in the resolutions.

References

Sergio Torres Aguilar, Xavier Tannier, and Pierre Chastang. 2016. Named entity recognition applied on a data base of medieval latin charters. the case of chartae burgundiae. In *3rd International Workshop on Computational History (HistoInformatics 2016).*

Alan Akbik, Tanja Bergmann, Duncan Blythe, Kashif Rasul, Stefan Schweter, and Roland Vollgraf. 2019. Flair: An easy-to-use framework for state-of-the-art nlp. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics (demonstrations)*, pages 54–59.

Alan Akbik, Duncan Blythe, and Roland Vollgraf. 2018. Contextual string embeddings for sequence labeling. In *Proceedings of the 27th international conference on computational linguistics*, pages 1638–1649.

Roman Bleier, Eva Ortlieb, and Florian Zeilinger. 2023. Der regensburger reichstag 1576—digital.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the association for computational linguistics*, 5:135–146.

Ger Dijkstra, Nienke Groskamp, Rik Hoekstra, Marijn Koolen, Esger Renkema, Ronald Sluijter, Frank Smit, and Joris Oddens. 2025. Entities recognised in the resolutions of the states general of the dutch republic (1576-1796).

Tomaž Erjavec, Maciej Ogrodniczuk, Petya Osenova, Nikola Ljubešić, Kiril Simov, Andrej Pančur, Michał Rudolf, Matyáš Kopp, Starkaður Barkarson, Steinþór Steingrímsson, et al. 2023. The parlamint corpora of parliamentary proceedings. *Language resources and evaluation*, 57(1):415–448.

Krzysztof Fokt and Maciej Mikuła. 2025. The digitalisation of the oldest legacy of the Polish and Polish-Lithuanian Seym in the framework of the IURA Project: dilemmas, limitations, prospects. In Florian Zeilinger, Roman Bleier, and Josef Leeb, editors, *Digitale Edition und vormoderner Parlementarismus/Digital Scholarly Edition and Pre-Modern*

Parliamentarism, pages 111–119. Vandenhoeck & Ruprecht.

Alex Graves, Santiago Fernández, Faustino Gomez, and Jürgen Schmidhuber. 2006. Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural 'networks. *ICML 2006 - Proceedings of the 23rd International Conference on Machine Learning*, 2006:369–376.

Rik Hoekstra, Marijn Koolen, Joris Oddens, and RGH Sluijter. 2025. Structure-derived incremental modeling: The case of the resolutions of the dutch states general. In Florian Zeilinger, Roman Bleier, and Josef Leeb, editors, *Digitale Edition und vormoderner Parlementarismus/Digital Scholarly Edition and Pre-Modern Parliamentarism*, pages 121–141. Vandenhoeck and Ruprecht.

Eero Hyvönen, Laura Sinikallio, Petri Leskinen, Senka Drobac, Rafael Leal, Matti La Mela, Jouni Tuominen, Henna Poikkimäki, and Heikki Rantala. 2025. Publishing and using parliamentary linked data on the semantic web: Parliamentsampo system for parliament of finland. *Semantic Web*, 16(1):SW–243683.

Stefan Klut, Rutger van Koert, and Ronald Sluijter. 2023. Laypa: a novel framework for applying segmentation networks to historical documents. In *Proceedings of the 7th International Workshop on Historical Document Imaging and Processing*, pages 67–72.

Marijn Koolen and Rik Hoekstra. 2022. Detecting formulaic language use in historical administrative corpora. In *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.

Marijn Koolen, Rik Hoekstra, Joris Oddens, and Ronald Sluijter. 2023a. Formulas and decision-making: the case of the states general of the dutch republic. *Proceedings http://ceur-ws. org ISSN*, 1613:0073.

Marijn Koolen, Rik Hoekstra, Joris Oddens, Ronald Sluijter, Rutger Van Koert, Gijsjan Brouwer, and Hennie Brugman. 2023b. The value of preexisting structures for digital access: Modelling the resolutions of the dutch states general. *ACM Journal on Computing and Cultural Heritage*, 16(1):1–24.

Marijn Koolen, Rik Hoekstra, Rutger van Koert, Ronald Sluijter, and Joris Oddens. 2025a. paragraphs of the resolutions of the states general of the dutch republic (1576-1796).

Marijn Koolen, Bas Leenknegt, Rik Hoekstra, Hayco de Jong, Sebastiaan van Daalen, and Esger Renkema. 2025b. Huygensing/republic-project: v1.0.0.

Marijn Koolen, Esger Renkema, Nienke Groskamp, Frank Smit, Jirsi Reinders, Ronald Sluijter, Rik Hoekstra, and Joris Oddens. 2024. Accessing the republic. entity extraction from the resolutions of the dutch states-general.

Marijn Koolen, Esger Renkema, Nienke Groskamp, Frank Smit, Jirsi Reinders, Ronald Sluijter, Rik Hoekstra, and Joris Oddens. 2025c. Accessing the republic. entity extraction from the resolutions of the dutch states-general. *DH Benelux Journal*.

Enrique Manjavacas and Lauren Fonteyn. 2022. Non-parametric word sense disambiguation for historical languages. In *Proceedings of the 2nd International Workshop on Natural Language Processing for Digital Humanities*, pages 123–134.

Ismail Prada Ziegler. 2024. What's in an entity? Exploring Nested Named Entity Recognition in the Historical Land Register of Basel (1400-1700).

Marie Puren, Fanny Lebreton, Aurélien Pellet, and Pierre Vernus. 2025. From parliamentary history to digital and computational history: a nlp-friendly tei model for historical parliamentary proceedings. *Digital Scholarship in the Humanities*, 40(Supplement_1):i75–i86.

Theodorus Helenus Franciscus Riemsdijk. 1885. *De griffie van hare hoog mogenden: bijdrage tot de skennis van het archief van de Staten-Generaal der Vereenigde Nederlanden*. M. Nijhoff.

Leanne Seaward and Stan Matwin. 2009. Intrinsic plagiarism detection using complexity analysis. In *Proc. SEPLN*, pages 56–61.

Ronald Sluijter, Rutger van Koert, Michael Baars, Marja Swüste, Michel van Gent, Esther van Gelder, Jesse Hollestelle, Ger Ruigrok, Ida Nijenhuis, and Joris Oddens. 2023. Republic pagexml ground truth handwritten resolutions states general.

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

Theo Thomassen. 2019b. *Onderzoeksgids: Instrumenten van de macht: de Staten-Generaal en hun archieven 1576-1796 (Band 2)*. Sidestone Press.

Rutger van Koert. 2023. Republic print dataset.

Rutger van Koert, Stefan Klut, Tim Koornstra, Martijn Maas, and Luke Peters. 2024. Loghi: An end-to-end framework for making historical documents machine-readable. In *International Conference on Document Analysis and Recognition*, pages 73–88. Springer.

Florian Zeilinger, Roman Bleier, and Josef Leeb, editors. 2025. *Digitale Edition und vormoderner Parlamentarismus/Digital Scholarly Edition and Pre-modern Parliamentarism: Eine interdisziplinäre Annäherung an frühneuzeitliche Quellen/An Interdisciplinary Approach to Early Modern Sources*, volume 114. Vandenhoeck & Ruprecht.