Textual repetition and variation in the Resolutions of the States General of the Dutch Republic

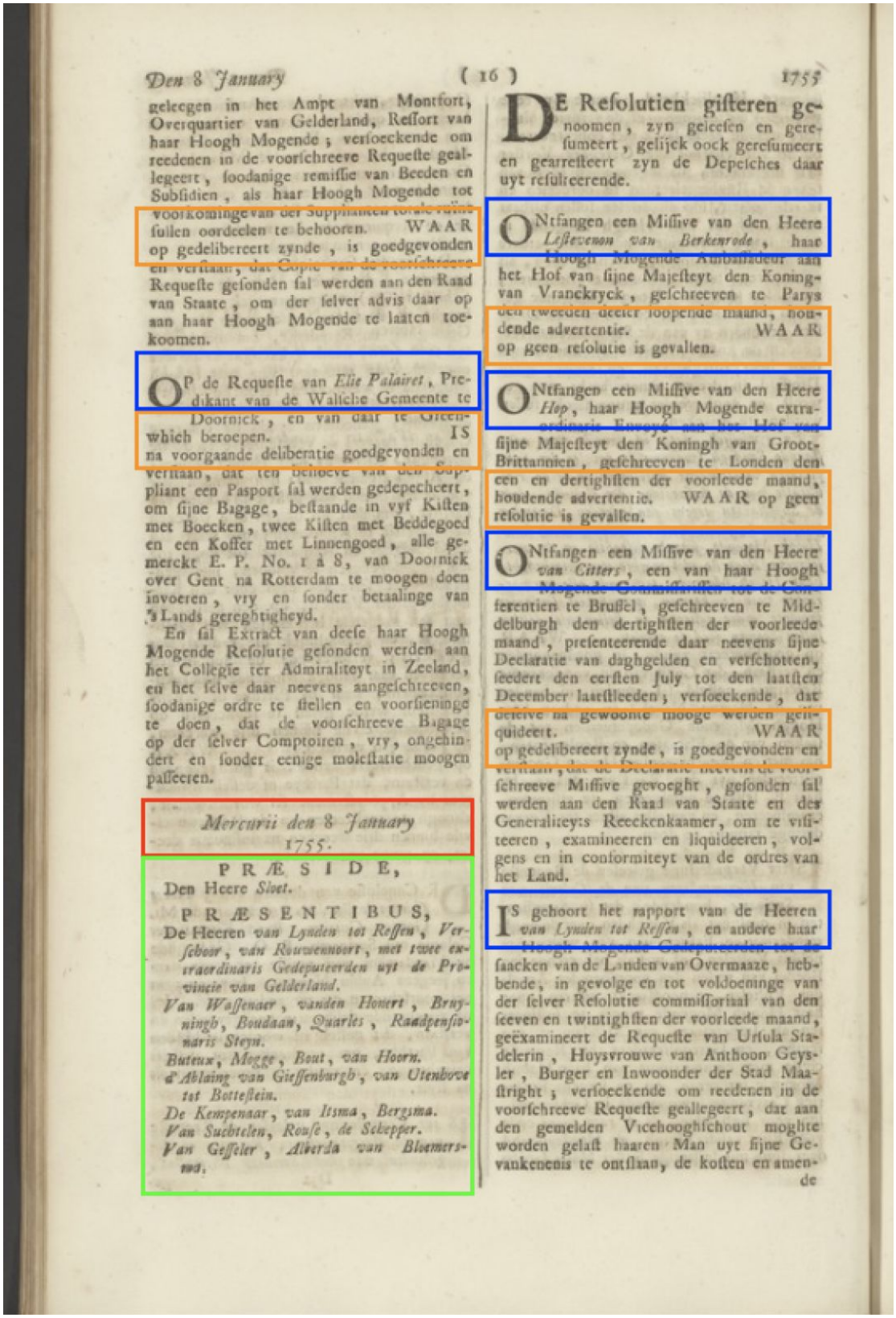
Marijn Koolen,[1] Rik Hoekstra,[1] Rutger van Koert,[1] Ida Nijenhuis,[2] Ronald Sluijter[2]
[1] KNAW Humanities Cluster
[2] Huygens ING

In the NWO REPUBLIC project, we are creating digital access to the corpus of the Resolutions of the States General of the Dutch Republic (1576-1796). This corpus contains the decisions made in the States General each day for a 220 year period. The resolutions were recorded using a standard structure and contain many standard formulations for aspects of the decision making process, including the source of the topic that was decided on (a formal request, a missive, etc.), whether a decision was reached and what that decision was.

This standardization and repetition of standard formulations can be exploited to alleviate some of the typical problems of digitizing large historical corpora. The effectiveness of NLP techniques to extract information like named entities, events, sentiments or topics is compromised by a combination of historical language variation, changes in conventions for using upper casing to signal important nouns and errors in the automated text recognition process (Traub et al. 2015, Mutuvi et al. 2018, Hill & Hengchen 2019, van Strien et al. 2020). Although there are ways to partially solve OCR or HTR (Handwritten Text Recognition) errors through post-correction (see e.g. Reynaert 2014, 2016), its impact on the quality of information extraction is variable. With high error rates, even post-correction models struggle to correct enough errors without introducing new ones. Human readers can often still recognize and read text that the above mentioned techniques cannot. Especially with some knowledge of the provenance and nature of the text, we can still make sense of it. If you know what textual phrases you are looking for, you can also use fuzzy string searching algorithms to identify them in low-quality OCR'ed and HTR'ed text.

The formulaic expressions in the resolutions makes them good candidates for fuzzy searching. In this paper we discuss the methods we developed[1] and the evaluation results of applying them on 100,000 pages and roughly 50 million words of resolution text. Although the task of extracting resolution elements has similarities with Named Entity Recognition (NER) and Detection (NED) there are important differences. First and foremost, the resolutions are not named entities, but textual summaries of the decision making process of the States General. The resolution summaries typically mention many named entities (persons, organisations, geographic locations and dates), but these are not the elements of interest in this task. Instead, we look for the textual elements that signal where a resolution summary starts, which part contains the decision that was reached, and where the resolution ends.

[1] See https://github.com/marijnkoolen/fuzzy-search for the fuzzy search library we developed.



Den 8 January (16) 1755

gelegen in het Ampt van Montfort, Overquartier van Gelderland, Reſſort van haar Hoogh Mogende; verſoeckende om redenen in de voorſchreeve Requeſte geallegeert, ſoodanige remiſſie van Beeden en Subſidien, als haar Hoogh Mogende tot voorkominge van der Supplianten totale ruine ſullen oordeelen te behooren. WAAR op gedelibereert zynde, is goedgevonden en verſtaan, dat Copie van de voorſchreeve Requeſte geſonden ſal werden aan den Raad van Staate, om der ſelver advis daar op aan haar Hoogh Mogende te laaten toekoomen.

OP de Requeſte van *Elie Palairet*, Predikant van de Waliche Gemeente te Doornick, en van daar te Greenwhich beroepen. IS na voorgaande deliberatie goedgevonden en verſtaan, dat ten behoeve van den Suppliant een Pasport ſal werden gedepecheert, om ſijne Bagage, beſtaande in vyf Kiſten met Boecken, twee Kiſten met Beddegoed en een Koffer met Linnengoed, alle gemerckt E. P. No. 1 à 8, van Doornick over Gent na Rotterdam te moogen doen invoeren, vry en ſonder betaalinge van 's Lands gereghtigheyd.

En ſal Extract van deeſe haar Hoogh Mogende Reſolutie geſonden werden aan het Collegie ter Admiraliteyt in Zeeland, en het ſelve daar neevens aangeſchreeven, ſoodanige ordre te ſtellen en voorſieninge te doen, dat de voorſchreeve Bagage op der ſelver Comptoiren, vry, ongehindert en ſonder eenige moleſtatie moogen paſſeeren.

Mercurii den 8 January 1755.

PRÆSIDE,

Den Heere *Sloet.*

PRÆSENTIBUS,

De Heeren *van Lynden tot Reſſen, Verſeboor, van Rouwenhorſt, met twee extraordinaris Gedeputeerden uyt de Provincie van Gelderland.*

Van Waſſenaer, vanden Honert, Bruyningh, Boudaan, Quarles, Raadpenſionaris Steyn.

Buteux, Mogge, Bout, van Hoorn.

d'Ablaing van Gieſſenburgh, van Utenhove tot Botteſtein.

De Kempenaar, van Iſma, Bergſma.

Van Suchtelen, Rouſe, de Schepper.

Van Geſſeler, Alverda van Bloemersma.

DE Reſolutien giſteren genoomen, zyn geleeſen en gereſumeert, gelijck oock gereſumeert en gearreſteert zyn de Depeſches daar uyt reſulteerende.

ONtfangen een Miſſive van den Heere *Leſtevenon van Berkenrode*, haar Hoogh Mogende Ambaſſadeur aan het Hof van ſijne Majeſteyt den Koning van Vranckryck, geſchreeven te Parys den tweeden deeſer loopende maand, houdende advertentie. WAAR op geen reſolutie is gevallen.

ONtfangen een Miſſive van den Heere *Hop*, haar Hoogh Mogende extraordinaris Envoyé aan het Hof van ſijne Majeſteyt den Koningh van Groot-Brittannien, geſchreeven te Londen den een en dertighſten der voorleede maand, houdende advertentie. WAAR op geen reſolutie is gevallen.

ONtfangen een Miſſive van den Heere *van Citters*, een van haar Hoogh Mogende Gecommitteerden tot de Conferentien te Bruſſel, geſchreeven te Middelburgh den dertighſten der voorleede maand, preſenteerende daar neevens ſijne Declaratie van daghgelden en verſchotten, ſeedert den eerſten July tot den laatſten December laatſtleeden; verſoeckende, dat deſelve na gewoonte mooge werden geliquideert. WAAR op gedelibereert zynde, is goedgevonden en verſtaan, dat de Declaratie neevens de voorſchreeve Miſſive gevoeght, geſonden ſal werden aan den Raad van Staate en des Generaliteyts Reeckenkaamer, om te viſiteeren, examineeren en liquideeren, volgens en in conformiteyt van de ordres van het Land.

IS gehoort het rapport van de Heeren *van Lynden tot Reſſen*, en andere haar Hoogh Mogende Gedeputeerden tot de ſaacken van de Landen van Overmaaze, hebbende, in gevolge en tot voldoeninge van der ſelver Reſolutie commiſſorial van den ſeeven en twintighſten der voorleede maand, geëxamineert de Requeſte van Urſula Stadelerin, Huysvrouwe van Anthoon Geysler, Burger en Inwoonder der Stad Maaſtright; verſoeckende om redenen in de voorſchreeve Requeſte geallegeert, dat aan den gemelden Vicehooghſchout moghte worden gelaſt haaren Man uyt ſijne Gevankenenis te ontſlaan, de koſten en amende

Figure 1. A page of resolutions from the printed volume of 1755.

We treat the extraction of information as a text collation problem (Gilbert 1973): the resolutions have textual overlap when it comes to the introduction of the resolution and its decision paragraph, with some unknown amount of textual variation (including variation in spelling, phrasing and text recognition errors), and the aim is to identify and align the textual repetitions and variations. In Figure 1, the meeting of Wednesday the 8th of January starts at the bottom left with a date template **<weekday> den <date>** (in this case *Mercurii den 8 January 1755*, highlighted in the red-colored box) followed by the president for that day (signaled by *PRAESIDE*) and the attendants (*PRAESENTIBUS*) highlighted in green.

The paragraphs 2-5 in the right column each represent a resolution, with formulaic openings (highlighted in blue in Figure 1):

- *Ontfangen een Missive van ...* ('Received a missive of …')
- *Is gehoort het rapport van …* ('Has been heared, the report of …')

The decision part is clearly signaled using extra whitespace before the capitalised word *WAAR* (highlighted in orange in Figure 1):

- *WAAR op geen resolutie is gevallen* ('On which no resolution was reached')
- *WAAR op gedelibereert zynde, …* ('On which has been deliberated …')

These phrases are part of a short list of expressions that are used frequently throughout the corpus. In the resolutions of 1705 alone, for the phrase '*Ontfangen een Missive van* ' the fuzzy search strategy finds 315 variations because of spelling differences and OCR errors.

The fuzzy searching algorithm takes as input a manually created phrase model, which is a list of keywords and phrases of interest, where each entry can have an optional list of alternative phrases or spellings, and uses character skip grams to find candidate strings in the text for each phrase in the model. The searcher can be configured with different thresholds for edit distance and length variations (candidates may be shorter or longer than the phrase in the model).

We used fuzzy string searching to identify formulaic expressions and iteratively built a corpus-specific phrase model with which we identify:

1. the date and attendance list of each meeting, which are followed by all the resolutions of that day,
2. resolution boundaries, e.g. where they start and stop in the running text, so we know which text belongs to which resolution,
3. different types of opening phrases that correspond to different types of sources (e.g. requests, missives, reports, etc., see Figure 2), and
4. the decision paragraphs that state what decision, if any, was reached.

We can add these elements as meaningful metadata labels to individual resolutions, which aids subsequent information access and analysis. Moreover, we can use the consistency of the structure of the resolutions to spot errors. For instance, we can check for cases where an opening formula was found but no decision formula. Or check for missing meeting dates by temporally ordering the ones we did find. In other words, the 'messy' output of the OCR process is turned into what Christoph Schöch (2013) calls 'smart data', i.e. cleaned, structured and semantically explicit layers of metadata.

Egense (2017) proposed to build word embedding models for improving search in low quality OCR. Although this would help certain aspects of information access (i.e. directed search using keywords), our approach has the advantage of offering a way to derive systematic metadata that can be used to enable faceted search and for systematic comparison of subsets of the corpus.

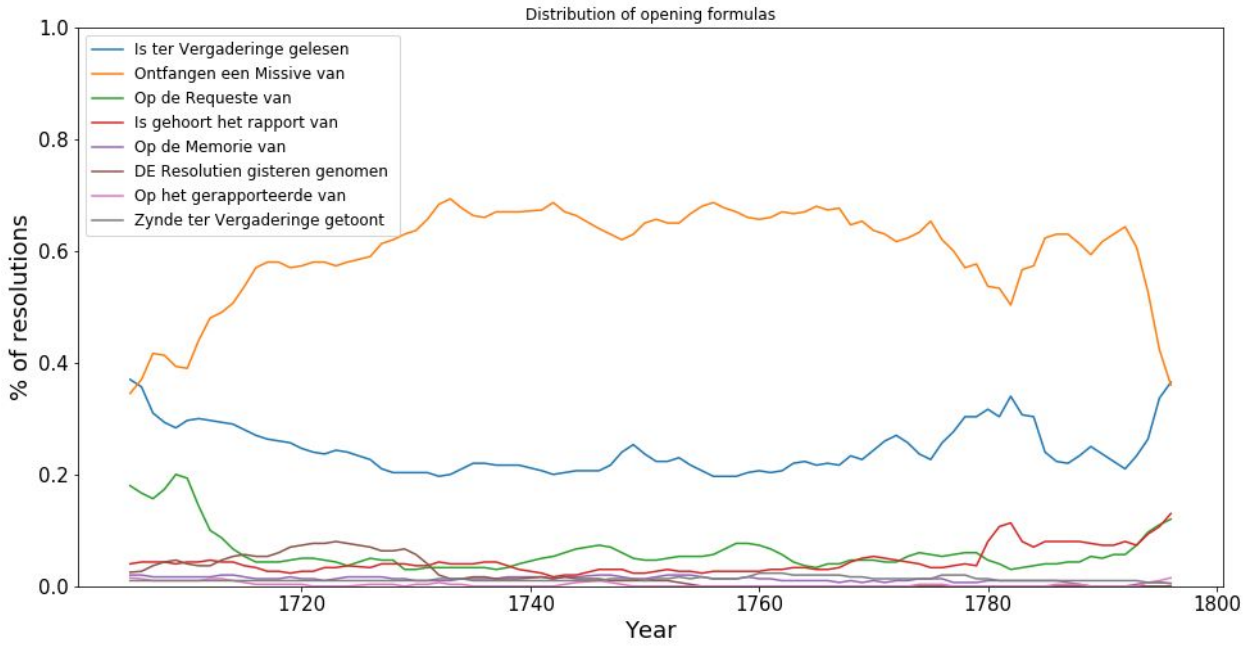


Figure 2. The relative frequencies of opening formulas for 189,895 resolutions. The percentages are smoothed using values of neighbouring years to reveal trends.

We built ground truth to evaluate the phrase model and the fuzzy searching and extraction process by randomly sampling historical dates, locating the pages containing the resolutions for those dates, and manually annotating the opening of the meeting, the attendance list and all resolutions openings, decisions and endings. For the meeting dates and attendance lists, our domain model and fuzzy searching approach reached a precision of 99% and recall of 93% on 300 randomly selected dates. For the resolution openings and decisions, we reached a precision of 90% and recall of 68%. We have not yet modelled insertions of extracts and letters, and therefore have not evaluated those aspects yet.

A qualitative analysis of the results shows that mistakes mainly fall in two categories:

- Non-standard openings: Most resolutions contain only one or two paragraphs, with a formulaic opening, a short decision paragraph and no formula for the ending, but which can be identified via the start of the next resolution. But a small fraction of the resolutions have no formulaic opening. Distributions like those in Figure 2 help us to identify periods with unknown opening formulas (i.e. formulas that are not in the model yet) or a lack of formulas, but there is no clear solution to this potential problem apart from manual annotation. Automated extraction struggles with such variation, thereby significantly reducing the value of the related metadata layer, making an access point that is biased towards periods of standardization.
- Not-recognized text: The OCR process occasionally misses some text on the page, leading to an incomplete textual representation of the image. In these cases, the fuzzy searching process needs to be able to deal with partial formulas.

One of the challenges is to find a systematic method for building our domain model of formulaic phrases, to ensure that changes in formulation over time or by different hands are captured, and that the set of phrases is complete or at least representative and covering the bulk of the material. Changes in formulas due to changes in spelling can be identified by using lower thresholds for what counts as a fuzzy match, and added as variants of a formula. But the bigger challenge is dealing with formulas that are missing or that are not standardized. For instance, there might be a period where the resolutions do not have a standard opening formula at all, and instead each is unique.

We have used early versions of this approach on other corpora and text genres. For instance, to start and end points of historical charters in digitized charter books, and to look up terms from back-of-book indexes in the same historical charter corpus and in the General Missives of the Dutch East India Company. Our approach is thus more broadly applicable, as formulaic expressions and genre templates were used in many other historic text documents, including other formal political documents, notary archives and newspapers.

References

Egense, T. (2017). Automated improvement of search in low quality OCR using Word2Vec. Digital Humanities in the Nordic Countries 2nd Conference. https://web.archive.org/web/20180613002357/http://dhn2017.eu/abstracts/#_Toc475332345

Gilbert, P. (1973). Automatic collation: A technique for medieval texts. *Computers and the Humanities*, *7*(3), 139-147.

Haentjens Dekker, R., Van Hulle, D., Middell, G., Neyt, V., & Van Zundert, J. (2015). Computer-supported collation of modern manuscripts: CollateX and the Beckett Digital Manuscript Project. *Digital Scholarship in the Humanities*, *30*(3), 452-470.

Hill, M. J., & Hengchen, S. (2019). Quantifying the impact of dirty OCR on historical text analysis: Eighteenth Century Collections Online as a case study. *Digital Scholarship in the Humanities*, *34*(4), 825-843.

Lopresti, D. (2008, January). Measuring the impact of character recognition errors on downstream text analysis. In *Document Recognition and Retrieval XV* (Vol. 6815, p. 68150G). International Society for Optics and Photonics.

Mutuvi, S., Doucet, A., Odeo, M., & Jatowt, A. (2018, November). Evaluating the impact of OCR errors on topic modeling. In *International Conference on Asian Digital Libraries* (pp. 3-14). Springer, Cham.

Reynaert, M. (2016, May). OCR Post-Correction Evaluation of Early Dutch Books Online-Revisited. In Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16) (pp. 967-974).

Reynaert, M. (2014, August). TICCLops:: Text-Induced Corpus Clean-up as online processing system. In Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: System Demonstrations (pp. 52-56). Dublin City University and Association for Computational Linguistics.

Schöch, C. (2013). Big? Smart? Clean? Messy? Data in the Humanities. Local Programs, Global Audiences, 2.

van Strien, D., Beelen, K., Coll Ardanuy, M., Hosseini, K., McGillivray, B., & Colavizza, G. (2020 Februari). Assessing the Impact of OCR Quality on Downstream NLP Tasks. ICAART 2020.

Traub, M. C., Van Ossenbruggen, J., & Hardman, L. (2015, September). Impact analysis of OCR quality on research tasks in digital archives. In *International Conference on Theory and Practice of Digital Libraries* (pp. 252-263). Springer, Cham.