BRILL

RESEARCH DATA JOURNAL FOR THE HUMANITIES AND SOCIAL SCIENCES 7 (2022) 1–10

BRILL OPEN

brill.com/rdj

Leiden University Resolutions Appendici Corpus (1575–1811)

Linguistics and Literature

Marten S. van der Meulen
Radboud Universiteit, Nijmegen, The Netherlands
marten.vandermeulen@ru.nl

Abstract

The Appendici to the Leiden University Resolutions of Curators and Mayors form a rich collection of documents from different genres, including letters, statutes and testimonies. Spanning the period between the founding of Leiden University in 1575 until its temporary dissolution in 1811, these documents are well-suited for historical linguistic research of Dutch in general, and for those interested in the relationships between norm and language in particular, as the period covers a key period in the codification of Dutch (ca. 1550–1804). In this data paper, the author introduces and describes the Leiden University Resolutions Appendici Corpus (LURAC), a single-domain (i.e., context of interaction), multi-genre diachronic corpus of 103,451 words, consisting of samples of the Appendici to the Leiden University Resolutions for ten time periods of 25 years between 1575 and 1811. Both raw data and metadata are available.

Keywords

Dutch – historical linguistics – language use – digital corpus – Leiden University

Online publication date: 12-2-2022

– Related data set "Leiden University Resolutions Appendici Corpus (LURAC)" with DOI www.doi.org/10.17605/OSF.IO/V9CT6 in repository "Open Science Framework"

1. Introduction

Leiden University, founded on February 8th, 1575 by decree of William of Orange, is the oldest university in the Netherlands and has been in continuous existence since the sixteenth century up to the present day. From the start, two bodies governed the university. The first was the Senate, which on paper consisted of a President (*Rector*), professors working at the university, and those doctors who lived and worked in the city of Leiden (although in practice especially this last group largely did not participate; cf. Sluijter, 2004, p. 22). The second governing body was the Board of Curators and Mayors, which consisted of three representatives of the States of Holland (the Curators) and the four mayors of the city of Leiden (Otterspeer, 2008, p. 31). Minutes, proceedings, and a plethora of other materials are available from both governing bodies for almost the entire life span of the university in the Leiden University Archives.

The Acta Senatus, or minutes of the Senate, were written entirely in Latin. By contrast, the Resolutions of Curators and Mayors were composed in Dutch from the founding of the university. These Resolutions consist of both minutes of the meetings of the Board, as composed by the Secretary of the Board, as well as documents that were discussed in the board meetings, including correspondence, budget proposals, declarations, and statutes. Until ca. 1750, these documents were copied in their entirety into the Resolutions, later only in-text reference was made to them. However, for the whole period, the originals were also produced as Appendici (*Bijlagen*). Both the Resolutions of Curators and Mayors and their corresponding Appendici are available in the Leiden University Archives (see the catalog by Hardenberg & Bouwman, 2006).

2. Context

Although the need for diachronic corpora, in the sense of "collection[s] of machine-readable, authentic texts, which [are] sampled to be representative of a particular language or language variety" (McEnery et al., 2010, p. 5), is well-established for Dutch (Coussé, 2010), such datasets are still in short supply for large parts of the history of Dutch (except for the oldest periods until ca. 1400, as represented in the Corpus Gysseling, the Corpus Van Reenen-Mulder and others, and a few recent examples, such as the Letters as Loot Corpus). Newly developed databases containing large amounts of data, such as Delpher (which contains over 120 million pages) and Nederlab (which consists of a variety of corpora and datasets totaling over 18 billion words), while extremely valuable for qualitative research within the historical sciences, are for the most

part unfit for most quantitative linguistic research, because of a lack of metadata, data curating and quality of optical character recognition (OCR; see Van der Sijs, 2019). However, recent years have seen efforts to increase the availability of diachronic corpora of Dutch, with initiatives such as the Language of Leiden Corpus (Assendelft, 2020) and the Corpus Historisch Nederlands (Lismont et al., 2019) arising as multi-genre corpora spanning the period 1500–1900, and C-Clamp, a large single-genre corpus spanning 1837–1999 (Piersoul, 2020). Unfortunately, at the moment these collections are not available to the wider research community, in part due to copyright issues, in part due to other issues. It is unclear whether or when these issues will be resolved. As such, the need for freely available diachronic corpora of Dutch remains high.

In light of this hiatus, the present data paper presents a curated diachronic corpus, consisting of samples of the Appendici to the Leiden University Resolutions. Aside from being one of the very few diachronic corpora of Dutch publicly available, the corpus has two distinct advantages. Firstly, while it contains different genres, all documents stem from the same social context of interaction or *domain*, namely the university. As such, the documents may be expected to be quite stable concerning style and register, which makes comparisons over a longer time period viable. Secondly, the proposed period, between the foundation of Leiden University in 1575 until the temporary disbandment of the university as an independent educational facility in 1811 (cf. Sluijter, 2004, pp. 14–15), closely mirrors the key period of selection and codification of the standardization of the Dutch language (cf. Haugen, 1966). The earliest attempts to set down the rules for Dutch date to the second half of the 16th century, with the first spelling book appearing in ca. 1550 and the first grammar published in 1584, and the first "official" codification of the language in 1804/1805 (Rutten, 2016). This close mirroring between available data and the development of the Dutch normative tradition makes the present corpus especially suited for questions about interactions between norms and language use. Additionally, the dataset can be used more generally for investigations into the historical development of the Dutch language or even for cross-linguistic questions, such as comparative investigations of the development of spelling systems.

3. Data Collection

3.1. *Selection of Texts*

Documents were collected from the original Appendici per period of 25 years between the foundation of Leiden University in 1575 until 1812, resulting in ten periods (1575–1599, 1600–1624, etcetera). As such, this corpus is more

fine-grained than the recently proposed diachronic corpora of Dutch, enabling different research questions. From each period, documents were selected that were written in Dutch; documents written in Latin, which were found especially frequently in the early centuries, or French, which became dominant over the course of the 18th century, were not included. As a complete collection was outside of the scope of the present project, we decided to sample ca. 10.000 words per period, each time starting from the first document.

To facilitate future expansions of the corpus, and to avoid biases (e.g., including only beginnings of texts could lead to an overrepresentation of welcome formulas), we included only complete documents. We tried to approximate the word limit as closely as possible, but the length and number of words of these documents varied; consequently, the exact number of words varies slightly per period. Our approach resulted in the exclusion of certain longer texts; however, there is no reason to assume that the language of such documents differs fundamentally from the shorter ones. Also, by including more different texts, we avoid both linguistic and content biases that would result from including, for example, only one text in a particular period. Thus, we can expect our sampling method to be an accurate representation of the language of the university domain for each period.

3.2. *Transcription*

Initially, documents were produced by photographing the original Appendici texts from their source in the Leiden University Library (see Figure 1). I processed these documents and transcribed a sample of the collected documents (ca. 10.000 words from different periods) using Transkribus (Kahle et al., 2017). I then checked my transcriptions against a source publication made in the early 20th century (Molhuysen, 1913–1924). As it turned out, Molhuysen had copied the original documents faithfully (rare mistakes, such as *geruiuneert* for *geruineert*, were corrected). Because of this level of accuracy and because re-transcribing all documents would be very time-consuming while yielding only marginal benefits, the decision was made to copy documents from the source publication, which was digitally available.[1] The only disadvantage was that Molhuysen sometimes left out the formulaic endings of documents. When this appeared to have happened, I went back to the original and transcribed the endings. In the few instances when this was not possible, due to the unavailability of the originals in the Leiden University Archives, I have added "Possible missing ending" to the metadata file.

1 Available from http://resources.huygens.knaw.nl/leidseuniversiteit.

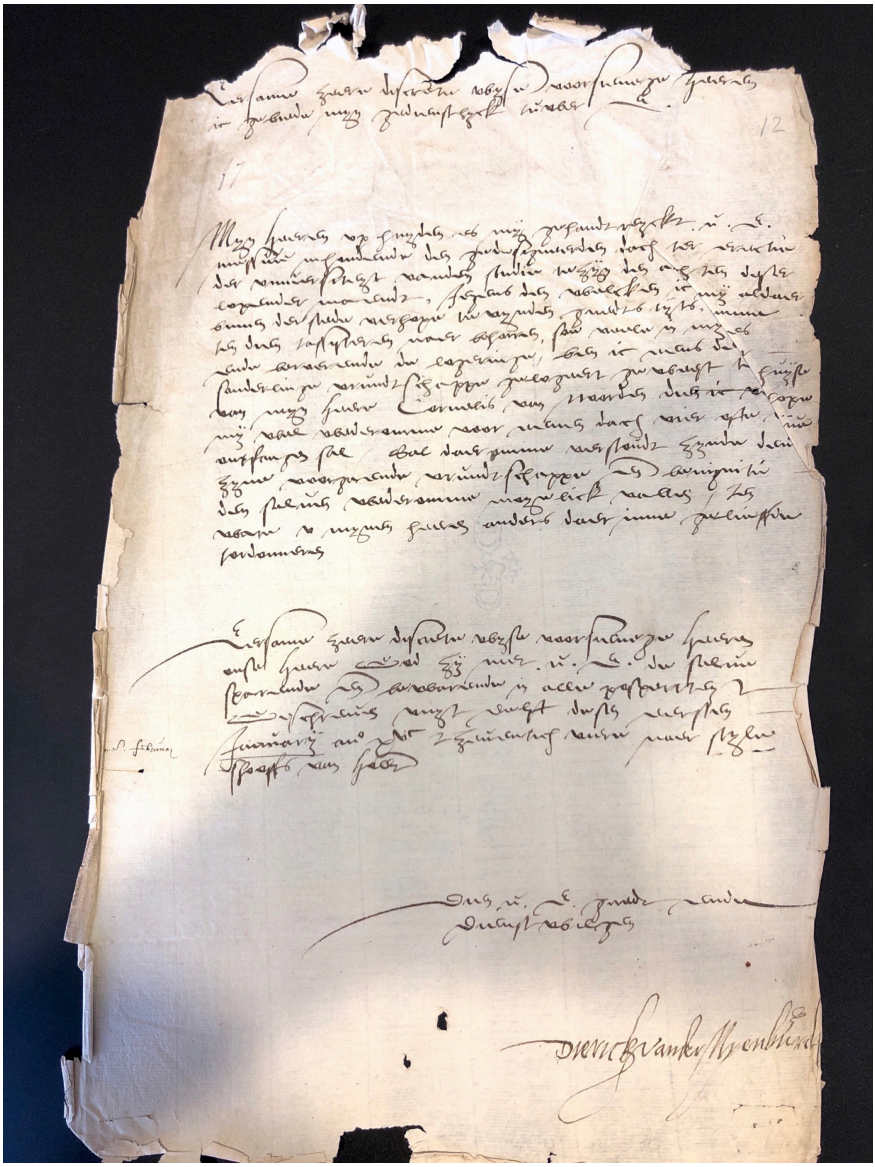

FIGURE 1 Example of material used from Leiden University archive (letter from D. Smaling and P. Vos to Board of Curators and Mayors, 31 January 1575)

My transcription of the Appendici is largely diplomatic, but I have made a few adjustments to ensure the suitability of the corpus for quantitative linguistic research questions:

– Word wrapping, which is employed in several different ways over the time period of the corpus (e.g. by using a space, colon, or hyphen) was resolved. However, hyphenation within a line, which was used for certain compounds, was retained.
– Abbreviations, including forms of address (e.g. *U. Ed.*) were retained. Superscript was resolved (e.g. *Burgerm.*^{en}) and dots were placed at the end of the abbreviation (e.g. *Burgermen.*).
– Spacing within a word, which was sometimes used for emphasis, was removed, because it made searching for these words impossible, and also because it substantially impacted word counts (e.g., *s e l f e* was transcribed as *selfe*).
– Line breaks were removed.

3.3. *Metadata*

Metadata was awarded to all documents for several parameters where available, following practices by, for example, Rutten & Van der Wal (2014). As an example, here is the metadata file for Appendix 2 from 1575:

– SOURCE: Archief van Curatoren, 1574–1815 - Universiteit Leiden (AC1 nr. 18 & nr. 38)
– TITLE: Voorstel van Sijne Excellentie of niet eenighe Collegien en Universiteit in den lande van Holland of Zeeland op te rechten.
– DATE (of composition): 2 januari 1575
– PERI: 1575–1599
– PLA (location of composition): Leiden
– DOMAIN: University
– GENRE: Correspondence
– WOCO (word count): 558
– WRI1 (writer/writers of the document): Paulus Buys
– GEND1 (gender of the writer/writers of the document): male
– PROF1 (profession of writer/writers of the document): lawyer

When data was not available for documents (which was especially likely for the writers), I filled in a question mark. To facilitate understanding, I made an exception for titles: when titles were not available from the source material, I followed Molhuysen's title attribution. When a document had multiple writers, the fields WRI1, GEND1 and PROF1 were repeated with a corresponding numeral (i.e., WRI2, GEND2 and PROF2). The domain was always University, but this was still added in order to compare the present data to other corpora.

4. Data Description

– **Leiden University Resolutions Appendici Corpus (LURAC) deposited at Open Science Framework – DOI:**www.doi.org/10.17605/OSF.IO/V9CT6
– **Temporal coverage:** 1575–1811

The data is available from the Open Science Framework (OSF) at https://www.osf.io/v9ct6/ (DOI 10.17605/OSF.IO/V9CT6). As Table 1 shows, the Leiden University Resolutions Appendici Corpus (henceforth LURAC) contains 165 documents covering the period between 1575 and 1811. The dataset consists of 103,451 words (average per document = 627 words); the longest document contains 4,746 words (Appendix 590; end of June 1631), the shortest contains only 62 words (Appendix 460; 17 April 1614).[2] For each document, both a raw. txt file and a metadata file are available.[3]

The distribution of different genres in LURAC is heavily skewed towards correspondence, both in the number of documents (77; 47% of the total) and the

TABLE 1 Years of origin, document and word token frequency per time period

Period	Years included	Number of documents	Word count
1575 - 1599	1575 - 1577	24	10,066
1600 - 1624	1600 - 1602, 1610, 1611, 1614, 1618, 1619	26	10,270
1625 - 1649	1625 - 1629, 1631	12	10,204
1650 - 1674	1650, 1651, 1653 - 1656	22	10,602
1675 - 1699	1675, 1676, 1678, 1686	13	10,106
1700 - 1724	1700, 1702 - 1705	13	10,193
1725 - 1749	1725 - 1730, 1732	18	10,436
1750 - 1774	1751, 1753, 1754, 1759, 1760, 1764, 1767 - 1769, 1771, 1772	16	10,323
1775 - 1799	1775, 1778, 1779, 1788	7	10,065
1800 - 1815	1801, 1802, 1804 - 1807, 1811	14	11,186
Total		165	103,451

2 All word frequency counts were produced using AntConc (Anthony, 2019). Other applications may produce slightly different numbers.

3 A lemmatized and Part-of-speech tagged version is highly desirable, but fell outside of the scope of the present project, because of the problems with spelling variation.

TABLE 2 Number of documents and word count per genre

Genre	Number of documents	Word token frequency
Correspondence	77	47,159
Declaration	10	5,086
Decree	10	7,394
Petition	9	4,065
Ordinance	7	5,514
Memorandum	6	4,889
Regulation	6	6,892
Resolution	5	1,840
Other	35	20,612
Total	165	103,451

number of words (46,797; 46% of the total). Although word token frequencies are somewhat better distributed among genres, the genres are not spread out evenly over the periods, and so a comparison between genres is, at present, not to be recommended. Table 2 gives an overview of documents and word token frequency per genre.[4]

For 100 of the documents, at least one writer is known; LURAC contains a total of 229 writers. This number is strongly skewed by Appendix 1023, which is a petition by 36 printers about the printing of disputations. This petition also influences the presence of different types of background in the corpus: including the writers of Appendix 1023, there are 42 printers in LURAC; however, they have only contributed to five different documents. Figure 2 shows the writers per type of background. As expected, the majority of writers worked in some capacity at Leiden University, mostly as secretaries to Curators and Mayors or as a professor, but also as a librarian and other capacities. The bar "Local government" contains, among others, pensionaries and aldermen. The bar "Supralocal government" contains five documents written by royalty, as well as several composed by the Grand Pensionary; professionals include notaries, a lawyer, an engineer, and others.

The writers in LURAC are overwhelmingly male: only two contributors are female. Appendix 462, which is the testimony of Judith Bays in the court case

4 A lemmatized and Part-of-speech tagged version is highly desirable, but fell outside of the scope of the present project, because of the problems with spelling variation.

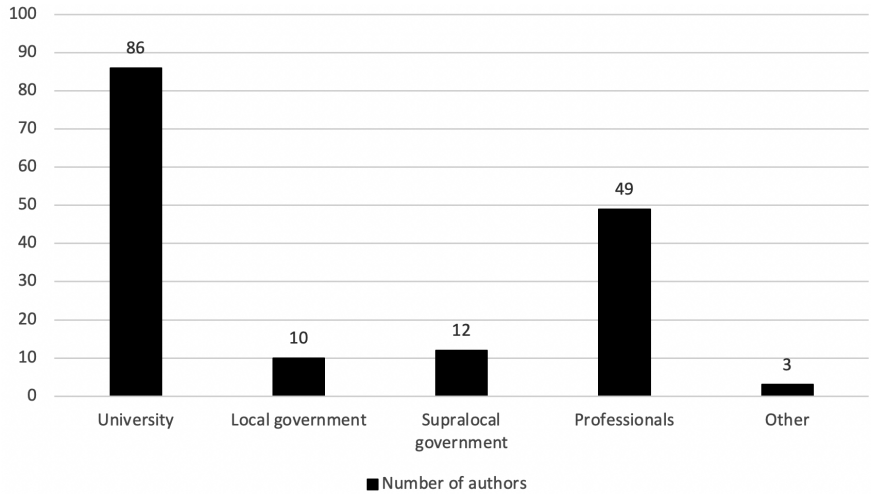

FIGURE 2 Writers (n=160) per type of background

of her son, Willem Merula, is the only document written solely by a woman; the aforementioned Appendix 1023 contains a Dieuwertje van der Boxe as one of the 36 signatories.

5. Concluding Remarks

We feel that the small but specialized Leiden University Resolutions Appendici Corpus presents researchers with a great resource to investigate different linguistic phenomena. Moreover, with this being one of the very few Open Access linguistic corpora of Dutch, LURAC presents an opportunity for interested non-specialists, teachers, and students of Dutch, not only for a look into the past of the language, but also to understand more about the production of (small) corpora. As such, it has value beyond our research community, as it can contribute to a better understanding of linguistic data and methods. We hope that the future will see the appearance of more such datasets.

Acknowledgements

This work was produced in the fall of 2020 during a Fellowship funded by the Maatschappij der Nederlandse Letterkunde. My thanks go to Gijsbert Rutten, Nicoline van der Sijs, Helen de Hoop, Mart van Duijn and Alan Moss for their input on this project.

References

Anthony, L. (2019). *AntConc* (3.5.8) [Computer software]. Waseda University. Available from https://www.laurenceanthony.net/software.

Assendelft, B. (2020). *De verfransing van het Nederlands: Een onderzoek naar de invloed van het Frans op het Nederlands tussen 1500 en 1900*. Grote Taaldag, Utrecht.

Coussé, E. (2010). Een digitaal compilatiecorpus historisch Nederlands. *Lexikos*, *20*, 123–142.

Hardenberg, H., & Bouwman, A. (2006). *Collection guide Leiden University Archives, Board of Governors, 1574–1815* (*ubl002*). Universiteitsbibliotheek Leiden. https://digitalcollections.universiteitleiden.nl/view/item/1887430.

Haugen, E. (1966). Dialect, language, nation. *American Anthropologist*, *68*(4), 922–935. https://www.doi.org/10.1525/aa.1966.68.4.02a00040.

Kahle, P., Colutto, S., Hackl, G., & Mühlberger, G. (2017). *Transkribus — A service platform for transcription, recognition, and retrieval of historical documents*. 19–24. https://www.doi.org/10.1109/ICDAR.2017.307.

Lismont, E., Van de Voorde, I., Rutten, G., & Vosters, R. (2019, December 13). *Spelling: Normen, gebruik en standaardisatie* (*16de tot 19de eeuw*). Colloquium Spelling in ontwikkeling, Gent.

McEnery, T., Xiao, R., & Tono, Y. (2010). *Corpus-based language studies: An advanced resource book*. Routledge.

Molhuysen, P. C. (1913–1924). *Bronnen tot de geschiedenis der Leidsche universiteit 1574–1811* (Vol. 1–7). Martinus Nijhoff. http://resources.huygens.knaw.nl/leidseuniversiteit.

Otterspeer, W. (2008). *Het bolwerk van de vrijheid. De Leidse universiteit in heden en verleden*. Leiden University Press.

Piersoul, J. (2020, October 10). *De compilatie van het Dutch C-CLAMP corpus* (*Dutch Corpus of Contemporary & Late Modern Periodicals*). [Online only]. Herfstvergadering van de Koninklijke Zuid-Nederlandse Maatschappij.

Rutten, G. (2016). Standardization and the myth of neutrality in language history. *International Journal of the Sociology of Language*, *2016*(242), 25–57. https://www.doi.org/10.1515/ijsl-2016-0032.

Rutten, G., & Van der Wal, M. (2014). *Letters as Loot. A sociolinguistic approach to seventeenth- and eighteenth-century Dutch*. John Benjamins Publishing Company.

Sluijter, R. (2004). "*Tot ciraet, vermeerderinge ende heerlyckmaeckinge der universiteyt". Bestuur, instellingen, personeel en financiën van de Leidse universiteit, 1575–1812*. Uitgeverij Verloren.

van der Sijs, N. (2019). Historische taalkunde en Digital Humanities: Samen naar een mooie toekomst. *Tijdschrift voor Nederlandse Taal- en Letterkunde*, *135*(4), 384–405.