The Kronieken Corpus: an Annotated Collection of Dutch/Flemish Chronicles from 1500–1850

Theo Dekker[1], **Erika Kuijpers**[2], **Alie Lassche**[1],
Carolina Lenarduzzi[1], **Roser Morante**[2] **and Judith Pollmann**[1]
[1]Faculty of History, Leiden University
[2]Faculty of Humanities, VU Amsterdam
{t.m.a.m.dekker,a.w.lassche,j.pollmann}@hum.leidenuniv.nl
{erika.kuijpers,r.morantevallejo}@vu.nl

Abstract

In this paper we present the Kronieken Corpus, a new digital collection of 204 local chronicles, containing almost 24 million words, written in Dutch/Flemish between 1500 and 1850. About half of these texts had not been published before. The manuscripts were photographed in 39 archives and libraries in The Netherlands and Belgium and subsequently transcribed and manually annotated by volunteers. The annotations include named entities and dates, as well as source mentions and attributions. The result is a unique, enriched historical corpus of original hand-written, non-canonical and non-fictional text by lay people from the early modern period.

1 Introduction

We present a newly transcribed and annotated dataset of local chronicles in Dutch from the period 1500-1850. The corpus has been compiled with the goal of developing a method to track and analyse the circulation, reception, evaluation and acceptance of old and new knowledge over time and across geographical locations by a lay public of mainly middle class authors. This work is part of the project *Chronicling novelty. New knowledge in the Netherlands, 1500-1850*[1] and the corpus is available for public use.[2] The historic period of 1500-1850 was chosen because it covers a number of societal changes that impacted knowledge production and circulation, such as the rise of the printing press, church reformations and the scientific revolution. This period also covered the so-called eighteenth century enlightenment and revolutionary era, as well as the political 'restoration' period of the early nineteenth century.

Local chronicles are chronologically organized accounts of events in the author's community. Following Pollmann (2016), who argued that historians of early modern Europe should more actively exploit the potential of the thousands of local chronicles that Europeans wrote between 1500-1850, we approached chronicles as collections of useful knowledge created by authors for future reference. Chroniclers collected information on a range of topics including local politics and history, crime, prices, public space and natural or cultural events that they deemed remarkable. Most of these texts were not written with a view to publication in print, but as manuscripts circulated among the literate middle classes of early modern towns and villages. Chronicles are one of the very few genres of narrative European texts that remained both ubiquitous and relatively stable throughout the early modern period. Therefore they can be used for comparative studies across both time and space about a wealth of topics. In the context of the research project Chronicling Novelty, we analysed what sources of information people considered reliable and how new information changed the way people reasoned over time (Dekker, 2022; Lassche and Morante, 2021; Lassche et al., 2022; Kuijpers, 2022). We believe that this dataset will be of unique value for research in history, digital humanities and historical linguistics, as well as for students of e.g. local politics, state formation, religious history, social conflict, and history of emotions.

After the discussion of some related work in Section 2, we describe the composition of the corpus in Section 3, the transcription process in Section 4, the annotations in Section 5 and the usage, distribution and maintenance in Section 6. Finally, we present a discussion in Section 7 and put forward some conclusions in Section 8.

2 Related Work

Digital historical texts in Dutch are made available by different institutions, and in different ways. Ex-

[1] https://www.nwo.nl/en/projects/vcgw17073.
[2] https://kronieken.transkribus.eu/.

Proceedings of LaTeCH-CLfL 2024, pages 243–252

amples can be found at the website of CLARIN,[3] the Huygens Institute[4] and the Institute for Dutch Lexicology.[5] Many texts can be searched and accessed via the Nederlab portal especially catering for historical linguists,[6] such as the Gysseling corpus, the Corpus Middelnederlands and the Corpus Oudnederlands. Most of these corpora consist of documents produced by institutions, such as the currently being digitized proceedings of the States General of the Dutch Republic (1576-1796).[7] Other resources contain Newspapers,[8] or the writings of important political or literary figures and scientists.[9]

In comparison to these existing corpora, this corpus is unique in several ways. It brings together a large set of **non-institutional** writings by a broad range of lay - often unknown - authors that are not archived in one place but scattered all over the Netherlands and Belgium. Similar to the collection of private letters confiscated from Dutch ships during the Anglo-Dutch Wars in the seventeenth and eighteenth century,[10] this collection represents the voices of individuals that belong to various social strata of society, who write on their own initiative and on topics that matter to them. Other than the seized correspondences, however, the chronicles are written in a larger geographical area comprising also current day Belgium and the Eastern and Southern inland provinces of the Netherlands (Rutten and Wal, 2011, 2014). We are not aware of a similar dataset in other languages.

3 Composition of the Corpus

When searching for chronicles that suited our goals we used the following selection criteria: First, we excluded family chronicles and regional chronicles – family chronicles lack the focus on public affairs, while regional chronicles were more often written for publication and by semi-professional historians. In order for our corpus to be searchable, we also decided we could only include texts that were (mainly) written in Dutch, even though French was also an important language in the Southern Low Countries, and there were also chronicles in Yiddish and Latin. Finally, we decided to focus on texts that were not only retrospective, but that also covered events in the (adult) lifetime of the authors, and were written contemporaneously.

The selection of the chronicles that would make up the corpus was carried out by the project leaders, both senior researchers in history, with the help and advice of student assistants, historians and archivists. The list of Chronicles that were selected can be found in our GitHub repository.[11] The collection process lasted from 2016 till 2018. After that time some more chronicles were identified and added.

98 local chronicles consisting of 131 volumes had been edited and published or transcribed for local archives or historical associations before. The DBNL,[12] an online database for literary texts in Dutch hosted by the Royal Library of the Netherlands, had already digitized some of these titles. The chronicles that were not already in the online database of DBNL, were newly digitized and added to it. The rest of the chronicles were manuscripts located in libraries and archives across Belgium and the Netherlands, or owned by private persons.

To find manuscripts we searched the digital inventories of the provincial archives in the Netherlands and Belgium for (variants of) words such as chronicle, annals, journal, history and diary. We did the same for local archives and a number of important libraries of which we knew or suspected that they could host chronicles. In this way we were able to add 106 unpublished chronicles (177 volumes) to our collection, that were sourced from 39 different archives and libraries and a few private collections. These archives and libraries had to be visited one by one, and every page of a chronicle manuscript had to be scanned. Some archives took on the task of scanning the chronicles themselves, but in most cases the ScanTent was used by the project team.[13] In combination with the DocScan app, the ScanTent enables the user to hold a document with both hands and scan it with their smartphone without pressing any button. DocScan

[3] https://www.clarin.eu/resource-families.

[4] https://www.huygens.knaw.nl/en/resources/.

[5] https://ivdnt.org/corpora-lexica/.

[6] https://www.nederlab.nl/onderzoeksportaal/?action=verkennen.

[7] https://republic.huygens.knaw.nl/.

[8] https://www.delpher.nl/nl/kranten.

[9] https://ckcc.huygens.knaw.nl/epistolarium/.

[10] https://brievenalsbuit.ivdnt.org/corpus-frontend/BaB/search/, https://prizepapers.huygens.knaw.nl/.

[11] https://github.com/chroniclingnovelty/chronicles-datasets.

[12] https://www.dbnl.org/.

[13] The ScanTent was developed as part of the READ project by members of the Computer Vision Lab of the Technical University Vienna and the Digitisation Preservation group of the University of Innsbruck. See https://readcoop.eu/scantent/.

automatically takes a picture once a page is turned.

The texts are all in Dutch/Flemish with sometimes quotations in other languages (mainly French or Latin). Spelling is very heterogeneous. Some texts, especially some sixteenth-century chronicles from the North-Eastern Netherlands have elements of the local dialect. Figure 1 shows a map of the Low Countries with the distribution of manuscripts over time and space.

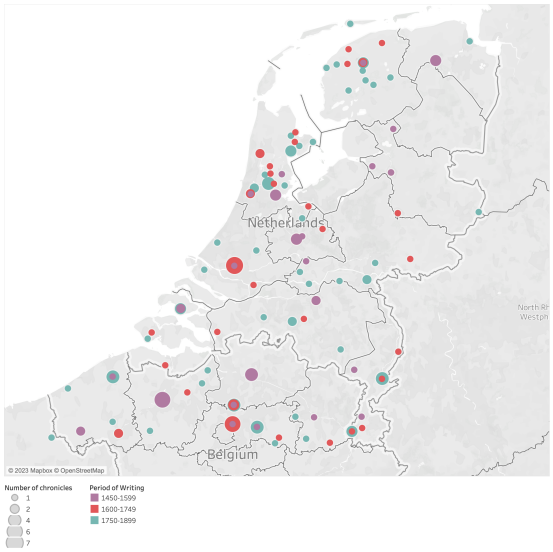

Figure 1: Map with number of chronicles per period and geographical points.

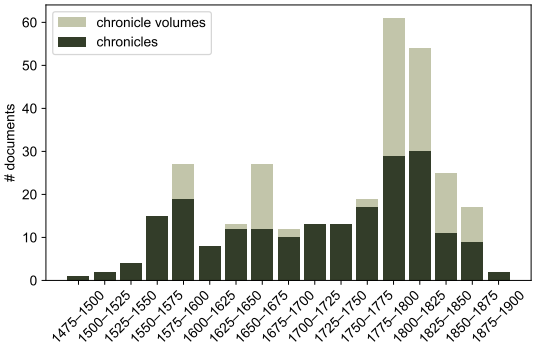

Figure 2: Distribution of chronicles in the Kronieken Corpus, visualized in bars of 25 years.

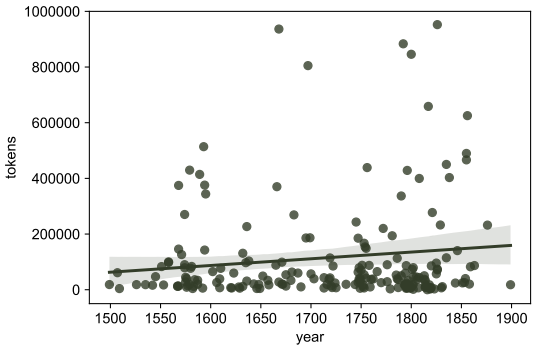

Figure 3: Number of tokens per chronicle in the Kronieken Corpus. For reasons of readability, one chronicle with 2.2 million tokens (written during many years until 1807) was excluded from this plot.

3.1 Units and size of corpus

Statistics about the full corpus can be found in Table 1. The total number of transcribed tokens is 23,871,380, belonging to 204 chronicles.

unit	amount
chronicles	204
chronicle volumes	308
tokens	23,871,380

Table 1: Size of the Kronieken Corpus.

In Figure 2, the distribution of the chronicles per time period is visualized in bars per 25 years. 1750 to 1800 is the period with more chronicles, whereas there are fewer for the first decades until 1525.

The scatter plot in Figure 3 shows the length of each chronicle in number of tokens. As can be observed, most chronicles contain less than 200,000 tokens, which applies to all time periods. The longest chronicles were written after 1650.

3.2 Data bias

Like any historical corpus there is both an institutional and a social bias in this corpus. Some categories of chronicles have had a better survival rate than others because of their content. Chronicles written in periods that were deemed important by later generations, such as the Dutch Revolt, or the Age of Revolutions have stood the test of time better than others. The fate of chronicles was also determined by the institutional context in which they were created. Thus, chronicles written by Catholic parish priests, who had no heirs, often remained in the parish. The same goes for chronicles written in convents. Town secretaries often passed on manuscripts to their successors, and in the course of our period some cities began to collect chronicles themselves. While most towns in the Low Countries had arrangements to keep their records safe, manuscripts that were written in villages may have been more vulnerable. Generally we may assume that many chronicles written by private individuals

may still remain in private collections while the majority got lost over time. Figure 4 shows the number of tokens dedicated to every year in the period 1500-1850, reflecting a bias in periods of war and upheaval. This graph is based on the 196 volumes that have date annotations allowing us to count the number of words on each year.

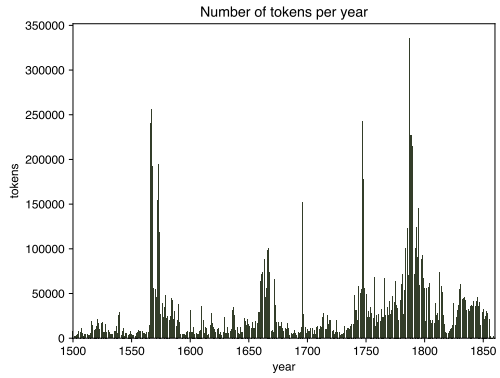

Figure 4: Number of tokens written per year based on the 139 annotated chronicles (196 volumes).

Based on the information that we collected about the authors[14] we could reconstruct the following social profile of the authors: Around 20% of the authors was anonymous, so all our knowledge about them comes from their texts. While the majority of the authors were men, we also identified 14 chronicles written by 16 female authors. 11 of them were nuns who wrote chronicles, probably in service of their convent. Although we had decided not to include convent chronicles, we made an exception for women's convents provided the chronicle focused on local rather than institutional events and resembled other local chronicles in style and content. Our assumption that chronicling was a typically urban activity seems to be correct. 79% of the chronicles in our corpus is written by urban dwellers, 21% in rural areas. For about 70% of the chroniclers we could establish their profession. 21% of all chroniclers were town secretaries, notaries, councillors, tax collectors or otherwise working in public administration. Another important group were clergymen, monastics, ministers and schoolmasters. All in all about half of the authors must therefore have had some form of (higher) education. However, over a quarter of the chroniclers earned their living in urban crafts and trades and almost 7% were farmers or farm hands. Both the real upper classes and nobility as well as unskilled labourers and the poor are underrepresented in this corpus (Kuijpers et al., 2024).

4 Transcription

Once the 106 unpublished chronicles (177 volumes) were photographed or scanned, they had to be transcribed. The scans of the chronicles were uploaded to a collection in Transkribus, a tool for handwritten text recognition (HTR) of historical documents by READ-COOP (Kahle et al., 2017). After uploading the scans, text regions and baselines were automatically detected. It turned out that in manuscripts with irregular hands or staining, text lines were missing or disrupted. In some cases the reading order of the lines was incorrect. Therefore, line segmentation had to be corrected manually by the team members and student assistants, which proved very time-consuming. After this, the scans were ready to be transcribed. This was done with the help of volunteers on the platform *VeleHanden* (ManyHands) which is being run by Picturae, a firm that specializes in the digitization of historical archives.[15] Picturae integrated the Transkribus web tool in the user interface of *VeleHanden*, allowing the volunteers to manually enter transcriptions that could later be used for training HTR models.

Every scan would be transcribed by one volunteer, and checked by another volunteer. On *VeleHanden*, these two roles were respectively the *invoerder* (transcriber) and *controleur* (controller). Every volunteer that was interested in the project could register as *invoerder*. The project team assigned specific volunteers (those who performed above average) the role of controller. Around 15,000 words of manually transcribed text were needed to train a HTR model that could be used to generate a transcription of the rest of the text. For the genre of a chronicle, this meant that about 40 pages of every handwriting needed to be transcribed by the volunteers. After training a model, its quality was evaluated using a test set. A model was considered good enough when the character error rate on the test set was below 4%. The automatically transcribed pages became visible for volunteers on *VeleHanden* to check. The *invoerder* now checked the transcription of the computer, and the controller performed a double check (Dekker et al., 2023).

To guide the volunteers in their work, transcrip-

[14] https://chroniclingnovelty.github.io/corpus-documentation/chronicles/.

[15] http://www.velehanden.nl.

tion guidelines were provided by the project team. These guidelines were based on the guidelines used by another *VeleHanden* project by the Amsterdam City Archive for the transcription of notarial deeds.[16] They contained transcription rules (about for example the use of capitals, punctuation marks, and illegible text), examples of often-used abbreviations in early modern written text, and lists of commonly used symbols and their meaning.[17]

The digitized texts of the other set of chronicles, those 98 chronicles that had been edited and published in the past, had to be uploaded to the same Transkribus collection as the manuscripts, to enable annotation. The digitized versions of these published chronicles contained all sorts of paratext, including introductions, footnotes, margin texts, and page numbers of the publication. All text that was not part of the original manuscript was removed. The varied manner in which the editorial additions to the chronicle were structured meant that most of this curation had to be done manually. Afterwards, the cleaned publications were converted to page XMLs, and uploaded to the Transkribus collection. Because the original page numbering was missing in these chronicles, we defined a page as a collection of 50 lines. However, since many of the chronicles lacked punctuation, some of these lines could turn out to be extremely long, while others were relatively short.

As a small team, with only five years' funding, we were unable to check and correct the transcriptions by the crowd by ourselves. Much of the proofreading was done by volunteers we selected and invited for that task. Even though the average quality of the transcriptions is good, the corpus is not consistent in the use of capitals, punctuation, and quite a few transcription mistakes remain. Most manuscripts until ca. 1710 were written in Gothic script, which would only be readable for a small group of experienced volunteers. Some handwritings are more difficult to read than others and also the condition of the volume, paper and ink could cause problems for even proficient transcribers. Missing or unreadable characters would be indicated by the transcribers with # or @ respectively.

[16] https://www.amsterdam.nl/stadsarchief/alleamsterdamseakten/. The project was named 'Crowd leert de computer lezen'.

[17] https://github.com/chroniclingnovelty/chronicles-datasets/tree/main/handleidingen/Invoerinstructie_Transkribus_Lite.pdf.

The transcription work started in July 2019, while the *Vele Handen* project closed by the end of 2022. Up until now a small group of volunteers is still transcribing directly in the Transkribus webtool. All but 5 volumes of the scanned manuscripts have been transcribed by December 2023.

5 Annotations

Because of the complexity of the task, annotation projects of historical corpora still make use of manual work by experts or volunteers although tools for automatic annotation are currently being developed (Tonelli and Menini, 2021; Arnoult et al., 2021; Sluijter et al.; Koolen et al., 2020; Koolen and Hoekstra, 2020). We performed the annotations with the Transkribus tool.

Once the chronicles had been transcribed we performed three annotation tasks. (i) We labeled named entities, dates, and page numbers. These labels should improve the searchability of the corpus for future users as well as enable our own analysis. Due to limited time, we were not able to annotate the full corpus. Instead, a subset of 139 chronicles (196 volumes) was annotated. (ii) In a smaller subset of 66 chronicles (85 volumes) the referencing to sources of information by the authors was annotated. (iii) In a third annotation project, attribution relations were tagged in another subset of the corpus (17 chronicles, 22 chronicle volumes). In the first task, we chose to annotate the corpus manually because our goal was to create the largest possible gold standard annotated data set. This differed from the annotation tasks 2 and 3, where the goal was to explore whether a limited set of manually annotated data would be sufficient to train computer models that would be able to automatically label source mentions and attribution relations. In the subsections below, the three annotation tasks are discussed in more detail.

units	**amount**
chronicles	139
chronicle volumes	196
tokens	12,709,875
`date`	172,974
`location`	292,726
`person name`	189,356

Table 2: Size of the subcorpus annotated with dates and named entities and number of annotations.

5.1 Annotation of dates, named entities, and layout features

Chronicles that were completely transcribed and controlled in the transcription project on *VeleHanden*, were made available in a second project on *VeleHanden*, in which volunteers annotated the chronicles. Guidelines were drawn up in which nine different labels were introduced and explained, accompanied by examples of text fragments in which the label had or did not have to be applied. Three content tags and six layout tags were determined: `date`, `location`, and `person name` were the content tags, and `pagenumber`, `margin text`, `lists and tables`, `copied text`, `image`, and `printed text` were the layout tags. The annotation guidelines are publicly available.[18] In Table 2, statistics can be found on the size of the subset, as well as the number of annotations of the three content labels (`date`, `location`, and `person name`).

The `date` tag contained an attribute, which meant that volunteers added the normalized date in an input field in the ISO 8601 format `yyyy-mm-dd`. This normalization step was essential since the chronicles showed a wide variety of ways in which dates were written. If volunteers were unsure about the normalized date (for example when a chronicler refered to 'St. Elizabeth's Eve'), they still tagged the text as `date`, but entered `xxxx-xx-xx` in the input field. They also used the `xx` if they were not sure about the exact day or month.

A mention of a land, region, place, street, water, or other known location or building, was tagged as `location`. If a location was conjugated to an adjective (for example 'a corps of Brandenburg troops'), the adjective was also tagged as `location`. The same was true for references to population groups, such as 'the Turcks' or 'the Venetians': they were also tagged as a `location`. The label `person name` was applied to mentions of a person's name. Titles of persons were tagged as well, and the same was true for professions, as long as they were accompanied by a person's name, such as 'Heer en Raed en advocaat Fiscaal Boreel' and 'Jan Stampijoen lantmeter'. The mention of a title or a profession only were not tagged.

The remaining six tags considered the layout features of the manuscript, rather than the content. If the author had used pagenumbers, this was annotated with the tag `pagenumber`. A reference to a folio number was also annotated, but when the page number or folio number was part of a reference '(see page X)', the label was not applied. The tag `margin text` was used when text was added in the margin or as a footnote. If a chronicler had noted information in a list or a table, for example the number of deaths per month, or price fluctuations, this was labeled as `lists and tables`. Text that was copied from another source and was recognized as such, was tagged as `copied text`. These fragments were in some chronicles indicated with quotation marks and/or a colon, in other chronicles words such as 'copy', 'extract' or 'resolution' were indications for a copied piece of text. Printed text, for example a pasted newspaper clipping, was tagged as `printed text`. Finally, if a scan contained an image, the label `image` was used.

Since the chance of errors was considered smaller in the annotation project than in the transcription project, the annotations were not double checked.

5.2 Annotation of sources

In order to get more insight into the reception of news and information by chroniclers, an annotation task was set up to label source mentions in chronicles (Lassche and Morante, 2021). A group of four volunteers, all having an above-average knowledge of the early modern Dutch language and culture, performed the task. They were provided with extensive guidelines in which source mentions were explained.[19] To extract source-related information, three labels were distinguished: `receiver`, the person receiving information; `source`, the instance bringing information to the receiver; and `perception`, how the source is bringing information to the receiver.

The label `perception` had four possible attributes: `oral/heard`, `written/read`, `seen`, or `else`. See the following examples, taken from the chronicles:

1. Deze morgen kwam <source> burgemeester Vorsterman </source> <receiver> ons </receiver> <perception: oral/heard> aanzeggen, dat wegens de ziekte, niemand in de kerk </perception>begraven mocht worden.

   This morning <source> mayor Vorsterman </source> came <perception: oral/heard>

[18] https://github.com/chroniclingnovelty/chronicles-datasets/tree/main/handleidingen/Annotatie_instructie_Vele_Handen.pdf.

[19] https://github.com/chroniclingnovelty/chronicles-datasets/tree/main/handleidingen/Annotatie_instructie_bronnen_Vele_Handen.pdf.

telling </perception> <receiver> us </receiver> that because of the disease, no one was allowed to be buried in the church.

2. 18 Februarij hebben <receiver> Wij </receiver> het Eerste in deze Stad in de <source> Amsterdammer Courant </source> van dien dag <perception: written/read> gezien </perception> dat Mevrouw Haere Koninglijke Hoogheijd Gemalin van de Heere Prince Erfstadhouder in 's Hage op den 16 dezer des Avonds te 11 Uuren Voorspoedig en Gelukkig was Verlost van een Gezonde en Welgeschapen Prins!

   On 18 February <receiver> we </receiver> have <perception: written/read> seen </perception> in the <source> Amsterdammer Courant </source> of that day that Her Royal Highness had given birth to a healthy and shapely Prince on the 16th at 11 in the evening in The Hague!

3. <receiver> Men </receiver> <perception: oral/heard> hoorde </perception> hoedat eenen boer sig zeer ongeluckiglijck verhangen hadt.

   <receiver> They </receiver> <perception: oral/heard> heard </perception> how a farmer had very miserably hanged himself.

Inter-annotator agreement (IAA) was calculated at two moments during the process of improving the guidelines, using the balanced F-measure (Hripcsak, 2005) (see Table 3). After the first calculation of the IAA, the F-scores were analysed. They showed that the guidelines caused the most confusion among the annotators regarding the label `source`. Annotators found it hard to distinguish between the description of an event ('Our Aldermen Court was heard') and the mention of a source ('We heard a strange rumour'). Guidelines were also not clear about self-references of a chronicler ('as I wrote on p. 23'). Some annotators interpreted this wrongly as a source mention.

	F-score 1		**F-score 2**	
	A1–A2	*A2–A1*	*A1–A2*	*A2–A1*
all	0.589	0.589	0.755	0.729
`source`	0.208	0.208	0.768	0.760
`receiver`	0.777	0.777	0.667	0.571
`perception`	0.707	0.707	0.754	0.699

Table 3: Inter-Annotator Agreement for the source annotation task in the first and second calculations.

The F-scores obtained in the second calculation of inter annotator agreement after improvement of the guidelines made it clear that much of the confusion was cleared up: especially the F-score of the label `source` was much higher than before, as shown in Table 3. Statistics on the size of the subset that was annotated, and the number of annotations that were made are in Table 4. An average of 93 sources were annotated per chronicle, compared to an average of 24 for receivers. The annotated data was used to train a classifier for automatic source annotation, but the low F-scores (below 0.4) of these models indicated a lack of success in this regard (Lassche and Morante, 2021; Lassche, 2024).

	amount
chronicles	66
chronicle volumes	85
`source`	6167
`receiver`	1597
`perception`	3391

Table 4: Size of the subcorpus annotated with sources and number of annotations.

5.3 Annotation of attribution

The extraction of attribution relations from text plays a relevant role in different NLP tasks such as the extraction of quotations and perspectives (Chen et al., 2019). An attribution relation (AR) is 'a relation ascribing the ownership of an attitude towards some linguistic material, i.e. the text itself, a portion of it or its semantic content, to an entity' (Pareti, 2012). An AR is typically expressed by three components: a *source*, a *cue*, and a *content*. A *source* is the entity that is the owner of the attributed abstract object, and can be a named entity, a noun or a pronoun. A *cue* is a lexical item which explicitly signals the ownership relationship between a source and a content. It is often a verb, but it can also be a noun, prepositional phrase, adjective or adverb. A *content* is a text portion which is perceived as meant to be attributed to the source. The following are examples of ARs:

1. <source> D'eene </source> <cue> gelooft </cue>, <content> dat ons Cristus suyvert van alle sonden </content>, d'ander heeft daertoe een vagevier gevonden.

   <source> Some </source> <cue> believe </cue> <content> that Christ purifies us from all sins </content>, others have found purgatory for this purpose.

2. <source>Een Heer, die destijds in Gecommitteerde Raden zat,</source> <cue> verhaalde </cue> mij eens, <content> dat hij driemaal bij den Hertog om audientie had laten vragen, zonder die te kunnen verkrijgen.</content>

   <source>A man, who was on the Committed Council at the time,</source> once <cue> told </cue> me, <content> that he had asked for an audience with the Duke three times, without getting it. </content>

One annotator who followed a training process labeled all the attribution relations in the chronicles under supervision of a senior researcher. Because the only existing guidelines for labeling attribution apply to contemporary English, guidelines that explained attribution relations in early modern Dutch texts had to be made. In Table 5, statistics can be found on the size of the subset that was annotated, and the number of annotations that were made.

	amount
chronicles	17
chronicle volumes	22
`source`	2880
`cue`	3546
`content`	3646

Table 5: Size of the subcorpus annotated with attribution and number of annotations.

During the process of improving the guidelines IAA was calculated for a sample of documents at two moments using the balanced F-measure[20] as shown in Table 6. Currently, experiments are run in which the manually annotated data is used to train a token classifier using BERT, as well as to let a generative model annotate more data.

	F-score 1		**F-score 2**	
	A1–A2	*A2–A1*	*A1–A2*	*A2–A1*
all	0.590	0.578	0.721	0.721
`source`	0.670	0.667	0.757	0.771
`cue`	0.624	0.601	0.812	0.801
`content`	0.503	0.497	0.570	0.574

Table 6: Inter-Annotator Agreement in the attribution annotation task.

6 Usage, Distribution and Maintenance

The corpus, including the transcriptions, meta-data, manual and automatic annotations and documentation has been made publicly available for future use under the creative commons license CC 4.0. All data is stored in a GitHub repository.[21] The digitized versions of published material are published on the DBNL website. The scans of the manuscripts that were uploaded in the Transkribus tool are accessible side by side with their transcriptions and annotations on the read and search website by READ-COOP.[22] Transcripts, annotations and images can be downloaded from this website by any user. Moreover at the 'back side' of this published collection it is still possible to correct transcripts if misreadings are found, and to add new scans of chronicles or missing transcriptions.

The options for future usage of the corpus are diverse. Chronicles belong to the type of material that are underrepresented in the digital resources for the humanities: original hand-written, non-canonical and non-fiction pre-modern material. However, they are considered of prime importance to historical linguists, and literary scholars as well as historians. Historical linguists are interested in chronicles because they give access to a historical linguistic variety that was 'filtered out' by professional printers, proofreaders and editors. For literary scholars, they offer vital access to reading and writing practices beyond the canonical authors. While medieval chronicles have been very well studied as a genre, and for the Netherlands have been digitally available for many years now, [23] early modern chronicles have only recently been rediscovered as an important resource. They provide a very valuable insight in the everyday experiences of life in historical urban and village communities.

The corpus of chronicles is also of great value for the digital humanities and computational linguistics communities. To begin with, corpora of this size and diversity of historic variants are very scarce, especially for Dutch. Such a corpus will allow us to make progress in processing historic variants of Dutch not only because it can be used to improve linguistic normalization tools, but also because it will allow users to train new tools. The additional layers of semantic annotation that will be provided with the corpus will allow the computational linguistics community to train new tools for the semantic processing of historical variants of Dutch. The corpus can be used for research purposes, as well as for teaching purposes. Students can be taught how to process this type of corpora with hands-on assignments. Finally, the corpus can be used to organize international shared tasks on processing historic variants of languages.

Finally, users should take into account that although we believe that chronicles are a genre of texts that have much in common, the diversity in size, topics, writing styles, motives and the pro-

[20] IAA was calculated between the annotator and one other expert.

[21] https://github.com/chroniclingnovelty/chronicles-datasets.

[22] https://kronieken.transkribus.eu/.

[23] http://www.narrative-sources.be/colofon_nl.php

ficiency of the authors make for a very heterogeneous corpus that sometimes hinders comparisons over time and across space.

7 Discussion

Our initial plan was not only to annotate a part of the corpus, but also to use the annotations to train machine learning systems to complete the annotations, for example to annotate source mentions and attribution relationships. However, this has proven to be more challenging than anticipated. Our exploratory experiments on automatically labeling source mentions demonstrated that the mentioning of sources showed so much variation and complexity that the training set was still too small, and the model used (CRF) was not the most powerful (Lassche and Morante, 2021; Lassche, 2024). In ongoing experiments aiming to annotate attribution relationships automatically, similar challenges are arising. Because the ways to automatically annotate data are rapidly expanding due to the swift developments in the field new avenues are opening for experimentation. We plan to train BERT classifiers (Devlin et al., 2019) and to annotate more data using generative models and appropriate prompting.

The corpus has several limitations. First, due to limited budget, time and staff, it was not possible to annotate the full corpus. We manage to annotate 197 out of 308 volumes, which amounts to 63% of the corpus. For the same reason, only 179 volumes have normalised date labels.[24] Second, some errors and misreadings remain. The transcription and annotation tasks were carried out by volunteers with varying proficiency in paleography and comprehension of historical language. The team of experts could not correct all transcriptions themselves but was assisted in this task by a selected group of volunteers. Third, apart from the earlier mentioned bias due to selection procedures by the team as well as the ravages of time, the following types of chronicles may be underrepresented in the corpus: chronicles written by women, chronicles written in rural areas, and chronicles written by lower class authors. Moreover, chronicles that are part of private collections, smaller archives, smaller towns and especially in archives that have not yet digitized their catalogues or inventories had a much smaller chance to be located by us.

[24] An overview with all chronicles and their annotation status can be found on https://github.com/chroniclingnovelty/chronicles-datasets/blob/main/handleidingen/Overview_Chronicles.xlsx.

8 Conclusions

We presented a corpus of 204 Dutch language chronicles from the period 1500-1850 counting almost 24 million words. The corpus has been transcribed manually by volunteers combined with automatic Hand Written Text Recognition as offered in the Transkribus Tool. The transcriptions have also been annotated by volunteers in three annotation tasks: A first general annotation of named entities, mentions of dates as well as elements in the lay out of the pages such as images, printed matter, tables and copied text. A second task focused on the annotation of sources of information mentioned by the author as well as the receiver of this information and the medium of communication, and a third task focused on attribution relations.

The result will be of value to both historians, students of historical literature as well as historical linguists. The additional layers of semantic annotation that are provided with the corpus will allow the computational linguistics community to train new tools for the semantic processing of historical variants of Dutch. Although a big effort was made to provide a quality resource, it was not possible to surmount some limitations posed by the magnitude of the project and the nature of textual data. In future work we will explore ways to surmount these limitations.

Acknowledgements

Research for this paper was funded through the NWO VC project 'Chronicling Novelty. New knowledge in the Netherlands, 1500-1850' directed by Judith Pollmann (Leiden University and Erika Kuijpers (VU Amsteram) and by the Network Institute of the VU Amsterdam, via the project "The Cycle of News in Chronicles from Eighteenth Century Holland: A Stylometric Approach".

References

Sophie I. Arnoult, Lodewijk Petram, and Piek Vossen. 2021. Batavia asked for advice. pretrained language models for named entity recognition in historical texts. In *Proceedings of the 5th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*, pages 21–30, Punta Cana, Dominican Republic (online). Association for Computational Linguistics.

Sihao Chen, Daniel Khashabi, Chris Callison-Burch, and Dan Roth. 2019. PerspectroScope: A window to the world of diverse perspectives. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 129–134, Florence, Italy. Association for Computational Linguistics.

Theo Dekker. 2022. God's Invisible Particles as an Explanation for the Rinderpest Outbreak (1713-1714): The Reception of Medical Knowledge in the Dutch Republic. *European journal for the history of medicine and health*, 79(1):152–168.

Theo Dekker, Erika Kuijpers, and Carolina Lenarduzzi. 2023. Van crowdsourcing naar echte burgerwetenschap. Investeer in de kwaliteit van samenwerking. *Stadsgeschiedenis*, 18(2):105–117.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

G. Hripcsak. 2005. Agreement, the F-Measure, and Reliability in Information Retrieval. *Journal of the American Medical Informatics Association*, 12(3):296–298.

P. Kahle, S. Colutto, H. Hackl, and H. Mühlberger. 2017. Transkribus - A Service Platform for Transcription, Recognition and Retrieval of Historical Documents. In *2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*, volume 04, pages 19–24.

Marijn Koolen and F.G. Hoekstra. 2020. The semantics of structure in large historical corpora. Digital Humanities 2020 : intersections, DH2020 ; Conference date: 20-07-2020 Through 25-07-2020.

Marijn Koolen, F.G. Hoekstra, I.J.A. Nijenhuis, R.G.H. Sluijter, Rutger Koert, van, Esther van Gelder, Gijsjan Brouwer, and H. Brugman. 2020. Modelling Resolutions of the Dutch States General for Digital Historical Research. Collect Connect: Archives and Collections in a Digital Age ; Conference date: 23-11-2020 Through 24-11-2020.

Erika Kuijpers. 2022. De informatiebronnen van Albert Louwen (1722-1798), kroniekschrijver te Purmerend. In Erika Kuijpers and Gerrit Verhoeven, editors, *Makelaars in kennis: Informatie verzamelen, verwerken en verspreiden in de vroegmoderne Nederlanden*, pages 131–158. Universitaire Pers Leuven.

Erika Kuijpers, Carolina Lenarduzzi, and Judith Pollmann. 2024. Profiling local chroniclers in the early modern Low Countries.

Alie Lassche. 2024. Information Dynamics in Low Countries' Chronicles (1500-1860). A Computational Approach.

Alie Lassche, Jan Kostkan, and Kristoffer Nielbo. 2022. Chronicling Crises: Event Detection in Early Modern Chronicles from the Low Countries. In *Proceedings of the Computational Humanities Research Conference 2022*, volume 3290 of *CEUR Workshop Proceedings*, pages 215–230. CEUR.

Alie Lassche and Roser Morante. 2021. The early Modern Dutch mediascape. detecting media mentions in chronicles using word embeddings and CRF. In *Proceedings of the 5th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*, pages 1–10, Punta Cana, Dominican Republic (online). Association for Computational Linguistics.

Silvia Pareti. 2012. A database of attribution relations. In *Proceedings of the Eighth Conference on International Language Resources and Evaluation (LREC'12)*, pages 3213–3217. ELRA.

Judith Pollmann. 2016. Archiving the Present and Chronicling for the Future in Early Modern Europe. *Past & Present*, 230(suppl 11):231–252.

Gijsbert Rutten and Marijke J. van der Wal. 2011. Local dialects, supralocal writing systems. The degree of orality of Dutch private letters from the seventeenth century. 14(2):251–274.

Gijsbert Rutten and Marijke J. van der Wal. 2014. *Letters as Loot: A sociolinguistic approach to seventeenth- and eighteenth-century Dutch*. John Benjamins.

Ronald Sluijter, Joris Oddens, Rik Hoekstra, Marijn Koolen, Rutger van Koert, Menzo Windhouwer, Hennie Brugman, and Femke Gordijn. Opening the Gates to the Dutch Republic: A Comparison between Analogue and Digital Editions of the Resolutions of the States General. In *Proceedings of the Digital Parliamentary Data in Action (DiPaDA 2022) Workshop*, volume 3133 of *CEUR Workshop Proceedings*, pages 158–166. CEUR. ISSN: 1613-0073.

Sara Tonelli and Stefano Menini. 2021. FrameNet-like Annotation of Olfactory Information in Texts. *Proceedings of the 5th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*.