

Identifying and Linking Names, Legal Statuses, and Punishments of the Dutch East India Company Convicts in Early Modern Batavia Using Supervised Named Entity Recognition (NER)

Muhammad Asyrafi and Fathiyarizq Mahendra Putra

Abstract

Under the Dutch East India Company (VOC) colonial law, enslaved and marginalised people were the most susceptible social group to judicial violence, such as corporal punishment and the death penalty. *Criminele Rollen* is one of the archival genres created by the VOC that recorded the information about it and has the potential to reveal the history of convicts, prisoners and death row inmates. As a genre of archives, *Criminele Rollen* has a distinct pattern of recording information. It was written chronologically and contains the names of the members of the Court of Justice (Raad van Justitie) who were present in court, the names of the accused and their legal status, the names of the prosecutors, information regarding cases, and court decisions. It also uses consistent phrases to indicate the sentences and other important information. The National Archives of the Netherlands holds roughly 236 bundles of *Criminele Rollen* sent to the VOC chamber in Zeeland. These archival documents are also part of the digitised archives provided by the De IJssberg Zichtbaar Maken project. The paper aims to uncover the names, legal statuses, and punishments of the slaves, convicts and death row inmates in *Criminele Rollen* using Named Entity Recognition (NER). We exercise EntityRuler and a custom pipeline in this study. By doing this, the article positioned itself as a preliminary attempt to uncover history from below using a digital humanities approach.

Keywords: Dutch East India Company (VOC), *Criminele Rollen*, enslaved people, death penalty, digital humanities, Named Entity Recognition (NER)

1. Introduction

In the early modern period, judicial violence, such as capital and corporal punishment, involved excruciatingly painful and brutal methods. This period was characterised by excessive violence, and scholars such as Michel Foucault and Pieter Spierenburg have extensively studied the violent punishments of this era.¹ However, these studies primarily concentrate on the European experiences of penal development. Meanwhile, the early modern period saw the rise of global networks and commerce, exemplified by the emergence of trading companies. These companies facilitated the spread of Western punitive practices worldwide, often mirroring the violent and painful punishments of Europe. The Dutch East India Company, or the *Verenigde Oost-Indische Compagnie* (VOC), was no exception. The VOC was granted octroi rights by the States General in 1602, which included the authority to exert punitive measures.² Addressing the lacunae in penal historiography, in recent years, the study on the VOC penal practice gained more attention.³ This trend offers a new insight into the history of penal practice. The history of penal practice in the early modern period is inseparable from the history of slavery. As a consequence, the study on the VOC penal practice also sheds light on enslaved populations and their struggle.

Operating under the influence of Roman-Dutch law, the VOC established a legal framework in its controlled territories, especially Batavia, which was the centre of the VOC operation in Asia. This system was based on Roman-Dutch legal principles and was crafted to uphold social order and reinforce control. In 1642, the VOC established the Batavian Statutes to regulate issues not addressed by Roman-Dutch law. The Batavian Statutes were applied to all inhabitants of Batavia, regardless of their ethnicity. Robert Cribb regarded the VOC's application of the Batavian Statutes in Batavia as the pinnacle of Dutch legal universalism throughout its colonial rule. Moreover, the application of the Batavian Statutes also reflected the European spirit of the age in Asia.⁴ Aside from the context of legal universalism, the penal practice in VOC-controlled Batavia mirrored the brutal practices of early modern Europe, including breaking on a cross or wheel, beheading, hanging, drowning,

¹ Antony, Carroll, and Pennock, *The Cambridge World History of Violence*; Foucault, *Discipline & Punish*; Spierenburg, *The Spectacle of Suffering*; Spierenburg, 'Judicial Violence in the Dutch Republic'; Bennett, *Capital Punishment and the Criminal Corpse in Scotland*; King, *Punishing the Criminal Corpse*.

² See Weststeijn, 'The VOC as a Company-State'.

³ Matthias Van Rossum argued that most of the works concerning the VOC's penal history focused on 'banishment' or 'exile'. See Rossum, 'The Dutch East India Company in Asia'; For other works on the VOC's penal history see Ward, *Networks of Empire Forced Migration in the Dutch East India Company*; Cribb, 'Convict Exile and Penal Settlement in Colonial Indonesia'.

⁴ Cribb, 'Legal Pluralism and Criminal Law in the Dutch Colonial Order'.

burning at the stake, mutilation, flogging, and branding.⁵ Additionally, the VOC established a judicial system with courts and legal officers who administered these severe punishments to enforce their authority.

The Statutes outlined the legal procedures that needed to be followed before the death penalty could be carried out. This included a formal trial by the Council of Justice (*Raad van Justitie*), where evidence was presented, and the accused had the opportunity to defend themselves. Based on the judge's decision, the council would then deliberate and issue a verdict.⁶ Although the Batavian Statutes formally disregarded ethnicity, in practice, the application of the death penalty in Batavia often depended on the legal status and social rank of the accused. Slaves and marginalised individuals were more likely to receive harsher punishments, including death, compared to free citizens or higher-ranking officials. This discriminatory aspect of the legal system reflected the broader social hierarchies and power dynamics of the colonial context. The Governor-General of the VOC had significant influence over the administration of justice, including the imposition of death penalties. In some cases, the Governor-General could grant clemency or commute a death sentence, although such instances were rare and typically depended on the political and social context.⁷

In some regions, the VOC employed exceptionally cruel methods even by the early modern standard, such as shoving convicts into cannons and shooting them against walls.⁸ Moreover, death did not mark the end of violence toward these individuals; post-mortem punishments were common. The bodies (and often body parts) of convicts were burned or displayed outside the castle walls to demonstrate the colonial government's power and serve as a deterrent to others.⁹ The harsh legal and punitive measures reflected the VOC's broader strategy of maintaining control and instilling fear among the population.

⁵ Asyrafi, 'A Distant Mirror'.

⁶ La Bree, *De Rechterlijke Organisatie En Rechtsbedeling Te Batavia in de XVIIIe Eeuw*; Ravensbergen, 'Courtrooms of Conflict'; Gaastra, *The Dutch East India Company Expansion and Decline*; Balk, Van Dijk, and Kortlang, *The Archives of the Dutch East India Company (VOC) and the Local Institutions in Batavia (Jakarta)*.

⁷ Ravensbergen, 'Gered van de koloniale galg koloniaal strafrecht en gratieverlening toegepast op 'Inlanders' en Chinezen in Nederlands-Indië 1819–1848'.

⁸ See for example the execution in Ceylon under Peter Vuyst in Samperman, *De Onregtveerdige Justitie, Uytgevoert Door Den Gouverneur Petrus Vuyst, Tot Ceylon Nevens Het Regtveerdig Vonnis En Regt, Aan Hem Gouverneur Gedaan, Door Den Achtbaren Raad van Justitie Des Casteels Batavia*; VOC, *Kort En Naauwkeurig Verhaal, van 't Leven En Opkomst van Den Heer En Mr. Petrus Vuyst*. Ultimately, Vuyst was arrested and executed in Batavia in 1732.

⁹ Peter Elberveld's gruesome execution and post-mortem punishment in 1722 is one of the infamous executions in Batavia. A monument erected in 1722 to warn people of Elberveld's crime is still standing today in Jakarta. See Horton, 'Pieter Elberveld'.

Batavia was the centre of the VOC's operation in Asia and was founded in 1619. It rapidly grew into a vital trading hub for the VOC, serving as a key node in the global spice trade. Considering Batavia's importance to the VOC, severe penal measures were intended, among other objectives, to keep the peace and order of Batavia. Meanwhile, the city of Batavia heavily relied on the labour of Asian slaves. These slaves, however big their contribution to the functioning of the city of Batavia was, were treated harshly. They were also subjected to an unfair legal system. The slave population in Batavia comprised individuals from various parts of Asia, including India, Ceylon (Sri Lanka), and the Indonesian archipelago. They were often captured during VOC military campaigns or purchased from slave markets. The VOC also exploited the local tradition of coerced labour. The majority of slaves in Batavia were personally owned.¹⁰ The living conditions for them were generally poor, with many subjected to physical abuse, severe punishment, and inadequate nutrition. Female slaves were often forced to work for prostitution.¹¹ Despite their harsh treatment, the enslaved population played a crucial role in the economic and social life of Batavia, working in households, agriculture, and various trades. Their labour was so essential to the VOC that they were subjected to involuntary migration and transported to any location where labour was needed. Hans Hägerdal described the slave community as 'constantly in flux due to high mortality and involuntary immigration'.¹² On top of that, the legal framework of the VOC offered minimal protection to slaves, and their status made them particularly vulnerable to exploitation and harsh disciplinary measures. Under the VOC rule, the marginalised group in Batavia, especially enslaved people, was susceptible to the gruesome penal practice mentioned above. Between 1729 and 1739, while comprising only 30% of the overall population in Batavia, enslaved individuals accounted for 50% of those who were subjected to capital punishment.¹³ The rich and detailed information on the practice of violent punishments is recorded in the VOC colonial archives, particularly the *Criminele Rollen*. Within these documents, the voices of the enslaved population await revelation. Why is it important to extract information from the archive using a new technique? By examining this particular archive using the newly available digital technique, we aim to unravel the hidden information within. The new digital technique also promised to reduce the time needed to analyse the archives.

¹⁰ Kanumoyoso, 'Beyond the City Wall Society and Economic Development in the Ommelanden of Batavia', p. 112; Nicolaas, 'Maritime (Im)Mobility', p. 55.

¹¹ Bosma and Raben, *Being Dutch in Dutch East Indies A History of Creolisation and Empire, 1500–1920*, p. 30.

¹² Hägerdal, 'A Slave Economy in the East Indies', p. 93.

¹³ Asyrafi, 'A Distant Mirror'.

2. New digital approach to uncover the Batavian penal history

The past decades witnessed a massive digitisation effort of archival sources.¹⁴ This effort has opened new research possibilities, and in recent years, the digital approach has been gaining more attention in the humanities, including in the field of history.¹⁵ Distant reading, a relatively new technique, is one of the few examples of this digital approach. Franco Moretti coined the term in 2000, initially for literary studies. It was met with both support and criticism.¹⁶ Support and criticism aside, recently, an interesting and promising approach has been proposed by Luthra et al., who have explored the possibility of using Automated Named Entity Recognition to uncover indexes of marginalised peoples in the VOC colonial archives.¹⁷ Our study took inspiration from it and used a similar approach to uncover the names, legal statuses, and punishments of the slaves, convicts and death row inmates in *Criminele Rollen* using Named Entity Recognition.

Named-entity recognition (henceforth NER) is a subtask of information extraction that seeks to locate and classify named entities mentioned in unstructured text into pre-defined categories such as person names, organisations, locations, medical codes, time expressions, quantities, monetary values, percentages, and so on. The goal of named entity recognition is to find all occurrences pertaining to a given class of entities.¹⁸ The application of NER is increasingly popular for identifying an entity or object in documents, one of which is historical data-based research; related research that Won and Murieta conducted provides a NER overview for identifying place names based on historical corpora.¹⁹ This research explains the purpose of exploring and evaluating modern tools that could be used to automate the identification and retrieval of geographic information from historical texts, particularly from the Republic of Letters (RoFL) datasets; another study explains using NER for Identified Biodiversity records from the past which uses Images of 20,000 herbarium specimens from the Museum of Human and Nature Activities,

¹⁴ For instance ‘Sejarah Nusantara’, n.d., <https://sejarah-nusantara.anri.go.id/>; ‘TANAP’, n.d., <http://tanap.net/>; ‘Delpher’, n.d., <https://www.delpher.nl/>; *Prize Papers Online*, n.d., <https://brill.com/view/package/ppo>; ‘Old Bailey Online’, n.d., <https://www.oldbaileyonline.org/>; ‘Digital Panopticon’, n.d., <https://www.digitalpanopticon.org/>; Liesbeth Keijser, ‘6000 Ground Truth of VOC and Notarial Deeds 3.000.000 HTR of VOC, WIC and Notarial Deeds’, January 2020, <https://doi.org/10.5281/ZENODO.6414086>.

¹⁵ This relatively new approach is often referred to as Digital Humanities.

¹⁶ Moretti, *Distant Reading*; Spivak, *Death of a Discipline*; Underwood, ‘Distant Reading and Recent Intellectual History’ p. 530; Moretti, ‘Franco Moretti’, pp. 686–689.

¹⁷ Mrinalini Luthra et al., ‘Unsilencing Colonial Archives via Automated Entity Recognition’.

¹⁸ Hakenberg, ‘Named Entity Recognition’, pp. 1487–1488.

¹⁹ Won, Murrieta-Flores, and Martins, ‘Ensemble Named Entity Recognition (NER)’.

Hyogo.²⁰ The process or methods are varied. In order to perform NER based on classification, we created a corpus of data labels used for test data and training data manually.²¹ Although both datasets are digitised, reprocessing them was required to optimally perform the NER tasks with the available tools for us to evaluate. This process consisted of cleaning the datasets and translating the original markup into textual content that can be taken as input by the NER tools.²²

This chapter aimed to create a NER dataset for historical archives based on the names of the convict, their legal status, the punishments, and the crimes they committed. We want to know how to implement machine learning to find these variables. To achieve its goal, this chapter is structured as follows: we describe the archival source used as the corpus in section 2, and then in section 3, we describe the preparation of data and the software we used. Next, in section 4, we discuss the evaluation of the research. In section 5, we conclude the study by describing the findings and results of the code.

3. Archival sources

The archival source selected to be experimented upon in this study is the *Criminele Rollen* series. The reason behind the selection is the availability of the codex and the importance of the corpus in the study of the history of punishment in the VOC Batavia. *Criminele Rollen* is a genre of legal document produced by the VOC. It is produced by the Council of Justice (*Raad van Justitie*). As a genre of archives, *Criminele Rollen* has a distinct pattern of recording information. *Criminele Rollen* starts immediately with the record. It has no register and contains the essence of the case, verdicts, and court advice. However, the definition of ‘essential’ depends on the case and the council member. *Criminele Rollen* ends on 30 August each year. *Criminele Rollen* has a set of rules for recording information on crimes and court trials. This form or pattern of writing is always obeyed by the officer in charge.

The cases recorded in the *Criminele Rollen* ranged from theft and murder to more complex crimes like treason and rebellion. The detailed nature of these records allows for a comprehensive understanding of the types of crimes that were prevalent and how they were dealt with by the VOC’s legal system. The structure of the case record beginning from the top of the page is as follows:

Some of the distinct features of *Criminele Rollen* are as follows: firstly, the case recording was arranged chronologically in each bundle. One bundle contains

²⁰ Takano, Cole, and Konagai, ‘Advancements in Data Extraction from Natural History Collections’.

²¹ Ibid.

²² Won, Murrieta-Flores, and Martins, ‘Ensemble Named Entity Recognition (NER)’.

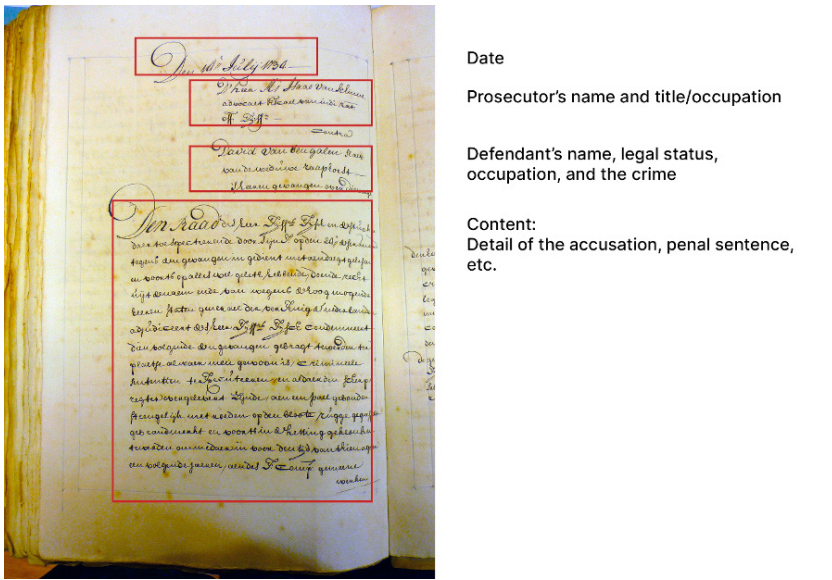


Figure 9.1. The structure of a Criminele Rollen page.

information about criminal cases for one administration year and disregards the importance of each case. For example, if a court meeting, trial, or appeal court was enacted in relation to a criminal case, this *Criminele Rollen* will record the ‘essential’ information about that hearing. If, after the hearing, whether it was on the same day or the next day, another case was tried, this other case would be written just under the previous record. If the verdict did not satisfy, or the judges did not come to a conclusion, and the case continued in the *Raad van Justitie* court, there would be another court.

Secondly, the *Criminele Rollen* was recording not only results but also the process. It means *Criminele Rollen* recorded the whole administration of justice in a full administrative year. This is different to other sub-genres of legal documents. For instance, *Processtukken* recorded in detail one particular case. *Criminele Rollen* recorded not only the verdict but also the process which preceded the verdict. This is important to any researcher who aims to understand the process of the court in the Court of Justice.

Third, the usage of the Latin language. It is a feature unique to legal documents. Particularly in this subgenre of legal document, the *Criminele Rollen*, terms such as ‘*Presentibus omnibus demptis*’ appeared often. This term means that all the members of the council were present except the name mentioned after ‘*demptis*’ or ‘*dempto*’. The name of the absent member then continued by the reason of their absence. The distinct traits of *Criminele Rollen* and the wealth of information it

contains render it a valuable asset for historical research, particularly in the field of penal history.

According to several studies, Named Entity Recognition (NER) for historical research faces challenges related to reference identification. These challenges include a) language spelling changes over time, b) spelling variation, c) OCR errors, d) the use of multiple languages, and e) general ambiguity in language use (Won, 2018, p. 2). Similarly, the study of *Criminele Rollen* encounters these issues along with additional problems: a) several cases of two-column documents, and b) the accuracy of the Handwritten Text Recognition (HTR) technology. These two additional challenges will be addressed in the following part.

4. Data and software

The process employed in this study is as follows: first, the corpus was selected based on its availability and relevance to the research topic. Next, the text data from the corpus was cleaned. After that, the labels to be extracted were determined. The cleaned corpus was then annotated based on these label entities. The annotated dataset was exported and used as the foundation for training the machine. Finally, the trained machine was tested, and the results were displayed using displaCy.

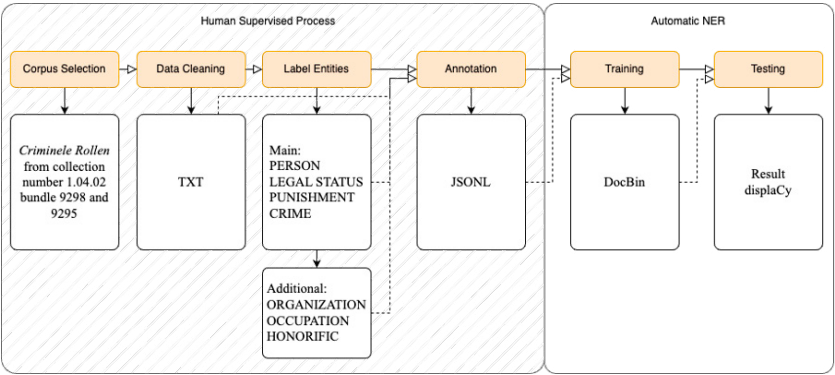
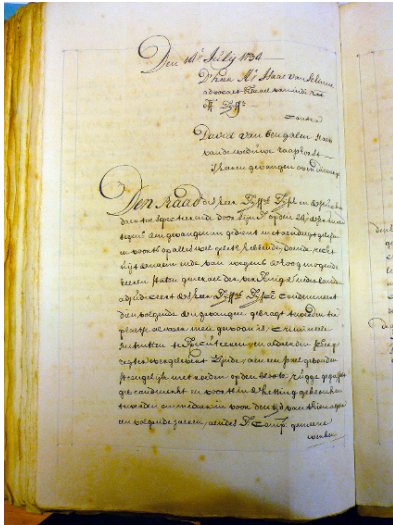


Figure 9.2. Workflow of the semi-supervised NER study for Criminele Rollen.

The data used for training and testing are taken from *Criminele Rollen*, which the Nationaal Archief keeps with collection no. NL-HaNA 1.04.02 bundle no. 9298 and 9295. In 2020, the project *De Ijsberg Zichtbaar Maken* (henceforth the *Ijsberg*) successfully digitised and processed a considerable portion of the VOC colonial archives. The project has also published and made the processed documents available to the

public.²³ *Criminele Rollen* documents that have been kept in the Zeeland chamber are included in this digitisation effort. The *Ijsberg* project utilised the Transkribus HTR+ model to turn the scanned and digitised archives into .txt files.²⁴ This study takes the .txt transcription files (henceforth the corpus) published by the *Ijsberg* project and processes them for Named Entity Recognition using supervised machine learning. An example of the corpus compared to the scanned document is below:



Den 14,e Julij 1734.
D' heer M^r Isaac van Schinne
advocaet fiscael van indi rat:
off: Eijss=r
contra
David van bengalen slaev
van de weduwe raaphorst
's heeren gevangen over dieverij
Den Raad des heer Eijss=r Eijsch en de stucken
daer toe sppecteerende door sijn E: op den 24. deser maend
tegens den gevangen in gedient niet aendaegt gelesen
en voorts op alles wel geleth hebbende, doende recht
uijt de naem ende van wegens de hoog mogende
heeren staten generael der verEenig de nederlanden
adjudiceert des heer Eijss=r Eijsch condemneert
dien volgende den gevangen gebragt tewerden ter
plaetse alwaer men gewoon is, crimineele
sententien te Executeeren, en aldaer den scherp,
regter overgelevert zijnde, aen een pael gebonden
strengelijc met roeden op den bloote rugge gegeesselt
gebrandmerkt en voorts in de ketting geklonken
tewerden ommedaer in voor den tijd van thien agter
een volgende jaeren, aen des E: Comp: gemeene
werken

Figure 9.3. A page of *Criminele Rollen* and the text transcription.

4.1. Corpus selection

This study selected 95 pages of *Criminele Rollen* from the collection number 1.04.02 bundle 9298 and 9295 as the training data. We acknowledge that the number of pages used here is underwhelming for a training dataset.²⁵ However, we expected this study to provide the preliminary basis for further research. The testing data is also taken from bundle no. 9298. We selected only the pages that mention corporal and capital punishment to highlight the penal sentence and legal status relation.

²³ Keijser, '6000 Ground Truth of VOC and Notarial Deeds 3.000.000 HTR of VOC, WIC and Notarial Deeds'.

²⁴ <https://readcoop.eu/model/dutch-handwriting-17th-19th-century/>

²⁵ Luthra et al.'s effort makes use of 2500 annotated pages. See Luthra et al., 'Unsilencing Colonial Archives via Automated Entity Recognition'.

4.2. Data cleaning

The corpus needs to be formatted before it is ready for NER, as the NER used in this study is a supervised machine-learning process that requires human supervision. In this study, data cleaning is performed manually. One of the challenges faced by the data cleaners is the format of the *Criminele Rollen* document. The Council of Justice's clerk sometimes wrote the document in two columns. While human readers can easily read the document in this format, the HTR engine reads it line by line, leading to improper transcription. This problem is evident in the following example:

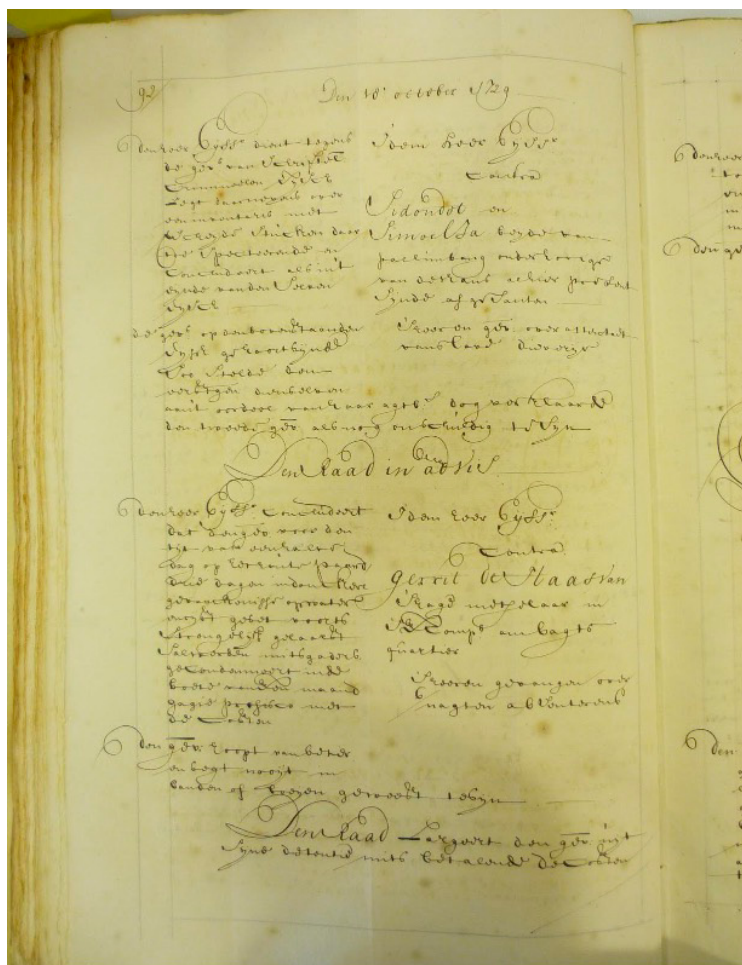


Figure 9.4. A page of *Criminele Rollen* with two columns format.

The transcription of the page provided by the *Ijsberg* project, as shown in figure 9.4, in a .txt format is as follows:

6
de
den
den
92
Den 18:' october 1729
den heer Eijss:r dient tegens
Idem heer bijss:r
de gev:s van schriftel:
Crimineelen Eijsch
Contra
zegt daarnevens over
een inventaris met
Sidondot en
sscheijde stucken daar
Timoelja beijde van
stde specteerende en
pallimbang onderhorige
Concendoert als in't
van dethans alhier present
eijnde van den Selven
Eijsch
sijnde afgesanten
gev:s op den bovenstaanden
Cheeren gev: over attentaet
Eijsch gehooft sijnde
vanslave dieverije
Soo stelde den
seerstgen: denselven
agtb:s dog verklaarde
aant oordeel van haar
den tweede gev: als nog onschuldig te zijn

in advis
Den Raad
heer bij Es.r concludeert
Idem heer bijss:

dat den gev. voor den
 tijt van een halve
 Contra
 dag op het houtte paard,
 Gerrit de Haasvan
 drie dagen indonckere
 Rhage metselaar in
 gevanckenisse opwater
 en rijst geset voorts
 'sE Comp:s ombagts
 strongelijk gelaarst
 quartier
 saltoerden mitsgaders
 gecondemneert inde
 Eheeren gevangen over
 boete van Een maand
 nagten absenterens
 gagie prosisco met
 de kosten
 gev: hoopt van beter
 en segt nooijt in
 —
 banden of boeijen geweest te zijn
 Den Raad Largeert den gev: uijt
 sijne detentie mits betalende de kosten

As clearly shown above, data cleaning necessitates human oversight, as the intricacies and nuances involved are beyond the current capabilities of the current state of the automated systems. Despite ongoing research and advancements in technology, no fully automated solution has been developed that can match the precision and judgment that human supervision provides in data cleaning.

4.3. *Label Entities*

Similar to the data cleaning process, the annotation process relies heavily on human supervision. The *Criminele Rollen* contains information on persons, legal status, crimes, and punishments. To maximise this information for penal history research purposes, we focused on four key entities: person, legal status, crime, and punishment.

4.3.1 Person

The label “person” encompasses any individual’s name that appears in the corpus. In this chapter, we annotated both European and Asian names. According to the writing system of the document, the convict’s name is always mentioned at the beginning of the record, followed by their legal status or occupation. The name then keeps being mentioned in the following passage, regardless of the legal status. This is dissimilar from the other archival genres, which put more importance on legal status rather than the name of an enslaved person. For example, according to Luthra et al.’s study, the enslaved person’s name is ‘hidden’ in the VOC testaments. The names are changed to simply ‘the slave’ as they were regarded as properties.²⁶ Such practice is not apparent in the *Criminele Rollen*.

It is important to note that the names of slaves were often recorded using names given to them by their slave owners, not the names they were born with. These given names, usually of European origin, were assigned because the slave owners found the slaves’ birth names difficult to pronounce. Thus, enslaved individuals were often recorded with names like Aurora, April, or Augustus. This practice highlights the erasure of their original identities in colonial records. Following the Dutch naming convention, the names of slaves were followed by their place of origin. Thus, typical slave names found in the *Criminele Rollen* include examples such as Aurora van Bugis, April van Malabar, and Augustus van Bali. In this study, we annotated both the name and the geographic surname together. Therefore, ‘Augustus van Bali’ is annotated as a single label entity, rather than just ‘Augustus.’ European names are annotated in a similar manner.

4.3.2. Legal status

The legal status is a crucial piece of information recorded in colonial archives. It helps to identify differences in the severity of punishments based on one’s legal status. For example, it allows us to determine if slaves were more prone to harsh punishments compared to free citizens. However, the legal status of a convict is not always mentioned in the *Criminele Rollen*. If the convict is an enslaved person, the term ‘*slaaf*’ or ‘*slavin*’ always follows. When the legal status is deemed unimportant, for example if a convict is a free person, the *Criminele Rollen* mentions the convict’s occupation instead. Additionally, terms like ‘*gevangen*’ (prisoner) and ‘*vrij*’ (free person) are considered legal statuses and are labelled accordingly in this study. An example from the corpus includes ‘*vrij javaan*’ or ‘*vrij balijers*,’ meaning free Javanese or free Balinese, respectively. It is important to note, as Bosma and Raben explain, that the term ‘free’ typically implied that an individual had previously

²⁶ Ibid.

been enslaved. In other words, being 'free' signified that the person was a manumitted slave.

4.3.3. *Crime*

The crime is always mentioned in the *Criminele Rollen*. In early modern Batavia, the judges followed Roman-Dutch law, which manifested in the Statutes of Batavia. The Statutes loosely listed a range of offences that were considered serious enough to warrant the death penalty. These included murder, treason, piracy, theft (especially repeated or aggravated theft), adultery, and certain acts of rebellion or insubordination against the VOC authorities. This particular label entity is straightforward in its writing and typically consists of a single word such as '*moord*' (murder) or '*dieverij*' (theft).

4.3.4. *Punishment*

The punishment is mentioned in the *Criminele Rollen*. The format of the word rarely changes. However, it is usually written in a long phrase which makes it difficult to label. An example of a punishment verdict in the training corpus is as follows:

alle de gevangens gebragt tewerden, ter plaatse alwaar meen gewoon is Crimineele sententien te Executeeren, en aldaar den Scherpregher overgelevert sijnde den gevangen mantaja met een ijsere pen van onderen Levendig gespit., op een paal geset, en alsoo ter dood gebragt, mitsgaders door de dienaars van de Iustitie so lange bewaert tewerden, tot dat sal sijn gestorven, wijders de gevangens Dormda eerst de regter hand afgehaept, en vervolgens aan een paal getrurgt tewerden dather de dood naavoegt Item den gev: Tjinar agterwaarts op een kruijs gebonden en van onderen op levendig gelede braakt tewerden, sonder slagen van gratie, mitsgaders in dien staat naahet buijten geregt gebragt aldaar onder opsigt en bewaringe der dienaren van de Iustitie op een rad geset, en na dat gestorven sal sijn niet en benevens deverdere geexecuteerde lichamen de vogelen des hemels ten sorcoije gelaten tewerden,

The text above discusses various aspects of punishment. There are several overlapping parts. For example, '*alle de gevangens gebragt tewerden, ter plaatse alwaar meen gewoon is Crimineele sententien te Executeeren*' describes the initial court decision to bring all convicts to the execution field.²⁷ Not all convicts received the same method of execution. For instance, in this text, capital punishment ranged

²⁷ The Dutch text can be freely translated as '[the council condemns] all prisoners to be brought to the place where it is customary to execute criminal sentences'.

from skewering, body mutilation, and breaking on a cross. The text for each verdict also contains multiple information about the punishment.

The first execution method is recorded as follows: *'en aldaar den Scherpregher overgelevert sijnde den gevangen Mantaja met een ijsere pen van onderen Levendig gespit., op een paal geset, en alsoo ter dood gebracht'*. It details the specific punishment for a convict named Mantaja. Mantaja was sentenced to be skewered from underneath with an iron quill, then set on a stake and brought to death. Labelling these entities can be complex. First, the entire initial part is labelled as 'PUNISHMENT.' Then, the phrase *'met een ijsere pen van onderen Levendig gespit'* is also labelled as 'PUNISHMENT.' Similarly, *'op een paal geset'* receives a 'PUNISHMENT' label, and *'ter dood gebracht'* is labelled as another 'PUNISHMENT.' Therefore, for the first convict we have three 'PUNISHMENT' labels.

The second convict, named Dormda, received different execution methods as described in this sentence: *'wijders de gevangens Dormda eerst de regter hand afgehapt, en vervolgens aan een paal getrurgt tewaterden dather de dood naavoegt'*. Firstly Dormda's right hand was cut off, indicated by the phrase *'eerst de regter hand afgehapt'*, which we label as 'PUNISHMENT'. Then the executioner had the bleeding Dormda to be hung on a stake until dead, described by the phrase *'aan een paal getrurgt tewaterden dather de dood naavoegt'*, and we label it as 'PUNISHMENT'.

The third convict, named Tjinan, received the following verdict: *'agterwaarts op een kruijs gebonden en van onderen op levendig gelede braakt tewaterden, sonder slagen van gratie, mitsgaders in dien staat naahet buijten geregt gebracht aldaar onder opsigt en bewaringe der dienaren van de Iustitie op een rad geset'*. Tjinan was broken on a cross as indicated by, *'agterwaarts op een kruijs gebonden en van onderen op levendig gelede braakt tewaterden'*, which we label as 'PUNISHMENT'. The phrase *'sonder slagen van gratie'*, that comes right after it, is also crucial as it refers to the method of execution. In breaking on a cross or wheel, there was a distinction between performing it with mercy and without mercy. *'Sonder slagen van gratie'* means without a mercy blow, indicating that the convict would be struck from below, starting from the legs and moving upwards, prolonging their pain. In the merciful version, the executioner would deliver a blow to the convict's head, causing death before the breaking of limbs. This is also labelled as 'PUNISHMENT'. The post-mortem punishment comes after that with the following text, *'naahet buijten geregt gebracht aldaar onder opsigt en bewaringe der dienaren van de Iustitie op een rad geset'*. The information in this part of the text contains the information telling us that the bodies of the three convicts were brought to outside of the castle and displayed. However, due to the text's length, only *'naahet buijten geregt gebracht'* is labelled as 'PUNISHMENT' to avoid complicating the machine learning process.

4.3.5. Additional label entities

In addition to these four label entities, we included organisation or ‘ORG’ and ‘HONORIFIC’ as label entities for further exploration using custom pipes. For example, honorifics such as ‘Mr.’ or ‘De Heer’ always precede the names of high-ranking VOC officials. In contrast, marginalised individuals, such as slaves or low-ranking soldiers who constitute the majority of convicts, do not have honorifics before their names. The ‘ORG’ label is used for entities like the Council of Justice (*Raad van Justitie*) and the VOC. These additional entities are beneficial in custom pipe settings; for instance, if a person’s name is preceded by an honorific, it can predict that the individual is a high official. Conversely, if the name is not preceded by an honorific, it is likely the name of a marginalised person. These two labels, however, are not the highlight of this study.

4.4. Data annotation

The corpora are annotated manually using Doccano based on the aforementioned label entities.²⁸ After annotation, the corpora are exported as a JSONL file. This JSONL file is then used to train and test the machine. The annotation dataset in the JSONL file is crucial as it serves as the foundation for the machine’s training and testing processes.

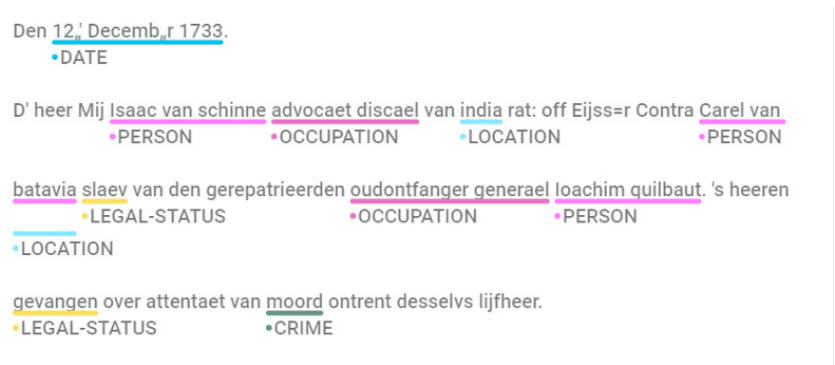


Figure 9.5. Annotation for Person, Legal Status, and Crime using Doccano.

As shown in Figure 9.5, the text contains errors because the corpora were produced using HTR technology. Due to the inaccuracies of HTR, some of the annotated words are slightly incorrect, as shown in Table 9.1:

²⁸ Nakayama et al., ‘Doccano’.

Table 9.1. Inaccuracies of the HTR-based transcription and the corrected labels

Words	Label	Correct Transcription
12,, ' Decemb,,r 1733	Date	12 December 1733
Isaac van schinne	Person	
Advocaet discael	Occupation	Advocaet Fiscael
India	Location	
Carel van Batavia	Person	
Batavia	Location	
Slaev	Legal-status	
Oudontfanger generael	Occupation	
Ioachim quilbaut	Person	
Gevangen	Legal-status	
Moord	Crime	

The corpora are annotated manually using Doccano based on the aforementioned label entities.²⁹ The annotated corpora are then exported as a JSONL file.

4.4.1. Model

This project utilised spaCy to process the corpus into a Recognised Entity or object. SpaCy is one of the libraries that has been developed in Python. It is an open-source Python library for NLP written in Python and Cython, and it offers pre-trained models for multi-language NER, including Dutch, as well as allowing training and deploying custom NER models on domain-specific corpora.³⁰

In this process, we trained several language-dependent NER models with distinctive configurations by using a) non-transformer spaCy small Dutch datasets and large Dutch datasets and b) the transformer model from SpaCy. All experiments were conducted through the command line interface using spaCy's train, evaluate and other commands. All models were trained on CPU and GPU using the Kaggle environment, and we trained 3 Dutch NER pre-trained models by spaCy with Doccano annotations. The training of the benchmark model was tried in Kaggle. The experiments started with the default model, which also serves as the baseline. We built the default configuration file in two steps: obtaining the base_config.cfg from spaCy's website as suggested and initialising the base configuration using the init command. The models used in this study are 'nl_core_news_sm',

²⁹ Ibid.

³⁰ Honnibal et al., 'SpaCy'.

`nl_core_news_md`, and `nl_core_news_lg`. For non-transformers, we are using ‘GroNLP/bert-base-dutch-cased’.³¹ After that, we visualise our model using `displaCy`. `DisplaCy` is a built-in dependency visualiser that lets us check the model’s predictions in a browser.³²

5. Results

5.1. Training

In the training phase, we process JSON data that can read the custom label, such as ‘HONORIFIC’, ‘ORG’, ‘LEGAL-STATUS’, ‘DATE’, ‘LAW’, ‘DURATION’, ‘OCCUPATION’, ‘PUNISHMENT’, ‘LOCATION’, ‘PERSON’, ‘CRIME’. This is necessary before it can be further processed in `spaCy`. All data files were converted into `spaCy`’s binary data format, a serialised `DocBin` before being used with the training command and other experiment management functions. The conversion was done using `spaCy`’s `convert` command through the command line interface and saved into `spaCy`’s `DocBin` file extension for testing method. In order to incorporate the `DocBin` during training, we use this command:

```
%%sh
python -m spacy train config.cfg --output ./ --paths.train ./training_data.spacy --paths.dev ./
training_data.spacy
```

This command initiates the training process using the configuration specified in the ‘`config.cfg`’ file. The output of this command consists of model files that include the words and labels provided earlier. These trained models will then be compared with unlabelled testing data in `.txt` file format during the testing phase. Additionally, it is important to note that during the training and testing phases, we will use two different `.txt` examples that have predefined labels. This allows us to evaluate the model’s performance comprehensively, ensuring that it can accurately identify and label data across different datasets. An example of training text contains the following passage:

Den 8:e Iulij 1730. W: m: Isaac Van Schinne advocaat fiscaal van india rat: aff: Eijsscher
Contra Jan Samuel Gunter van Sabelits in saven borger en Orenfried Ran uijt Lwickauw in
Sapen Corporael ten dienste der E Comp.s en Laast bescheijden geweest aan de nieuwpoot

³¹ Vries et al., ‘BERTje: A Dutch BERT Model’.

³² <https://demos.explosion.ai/displacy>

deser stad 'sheeren gev:e beijde over morshandel in scomp:s specerijen Den Raad als nu disponeerende op des heer Eijss.rs Eijsch admortem in dato 10 mejj Laastleden opende Iegens de gev: gedaen ende genomen Doende Regt, condemneert de gev:e beijde gebragt tewerden ter plaatse, alwaar men gewoon is Crimineele Sententien te Executeeren, en aldaar den Scherpregher overgelevert sijnde, den eersten gevangen met de roorde om den hals aande galge gestraft te werden, datt er de dood naavolgt, en den tweeden gev: aan een paal gebonden en strengelijk met roeden gegeeffelt tewerden, band voorts hem gev: voor al sijn leven uijt alle steden Corten en plaatsen onder de Iuristictie van dâc™ EComp: gehoorende sonder ooit daer weder in te mogen komen op pauze vanswaarder die Straffe Confisqueert voorts de helfte van alle de goederen en gelden vanden eersten gev: om deselve auso verdeelt te werden ontsegt den heer Eijss.r sijnen verderen Eijsch en condemneert de gev: daar en boven in de costen en misen van Justitie Aldus gearresteert n gesententieert op donderdag den 25: mejj 1730. bij den Edelen heere m:r Hendrik van Baarle Raad Extraordinaris van nederlands india en president, Item de heeren m:r Jsaak massis m:r gerardus gerlag, m:r gualter schutten, ne Jacob van den 6osch m:r Ian Blaankamer m:r Petrus Reijniers Thoren, m:r Cornelis van gaasbeecq, m:r Ian Andolph Sappius, en m:r Jacob Lakeman raden van Justitie, mitsg: gepromuntieert ende geExecuteert ten dage ende ter presentie als in den hoofde deses.

Furthermore, the training phase includes visualising the predicted entity labels using spaCy. Visualising the results helps in understanding how well the model is performing and in identifying areas where improvements might be needed. By employing spaCy's visualisation tools, such as displacy, we can create clear and intuitive visual representations of the labelled entities within the text. These visualisations provide insights into the accuracy and effectiveness of the model, making it easier to interpret the outcomes and refine the model accordingly.

As shown in Figure 9.6, the training process was conducted over multiple iterations, with key metrics recorded at various checkpoints. These metrics include Loss for 'tok2vec' and 'ner' components, as well as Entity F-score (ENTS_F), Precision (ENTS_P), Recall (ENTS_R), and overall Score. Initially, the model started with a high loss in both 'tok2vec' and 'ner', and an ENTS_F score of 0.00. This is typical at the beginning of the training process. However, the model showed significant improvement as training progressed, with losses dropping considerably and the ENTS_F score reaching as high as 99.87. This improvement indicates the model's increasing proficiency in identifying and labelling entities correctly.

Throughout the training, the model maintained high accuracy, with ENTS_F scores consistently close to 100. Although the final iterations showed some fluctuations in loss, the model overall maintained high precision and recall. The evaluation rate ultimately reached a peak of 1.00, demonstrating the model's robust performance. The learning score has reached a functional level, indicating that

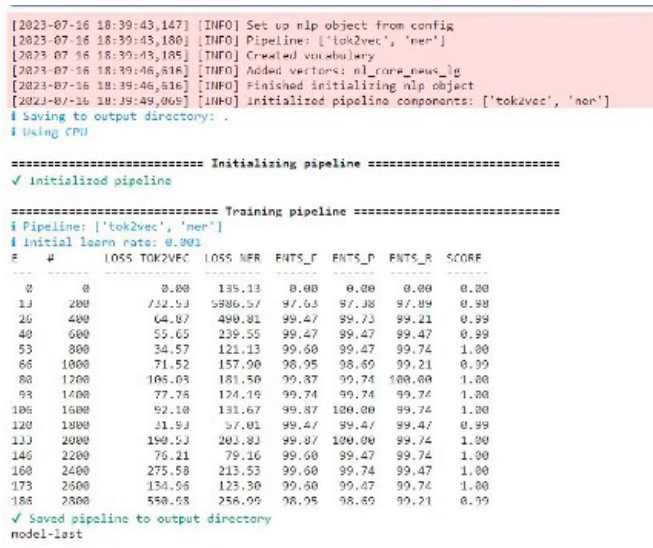


Figure 9.6. Evaluation Rate of the Training for spaCy `nl_core_news_lg` Model.

the model is performing well. With the training completed and the results deemed satisfactory, the model is now ready for testing and further validation. Future steps should focus on addressing any remaining issues with custom entity recognition and expanding the training dataset to ensure the model's continued accuracy and effectiveness.

5.2. Testing

A combined passage from bundle 9298 pages is used for the testing text. In this testing phase, we are using custom labels prepared in the previous stage, which are: 'LEGAL STATUS', 'OCCUPATION', 'PERSON', 'CRIME', and 'PUNISHMENT' alongside pre-defined labels from the model and additional custom labels, such as 'HONORIFIC,' 'ORG,' 'DATE,' and 'CARDINAL.' The testing process utilises the 'nl_core_news_lg' model, which is processed using Python in the Kaggle environment. The testing text includes the following passage:

Idem Heer Eijscher contra Martha Bintang van amboina wed:e van den inlands burger Matthijs Lawalatoe. 'sheeren gevange over attentaat van moord ent willen helpen ver-geven van den gerepatrieerden oud ontfanger generael Ioachim quilbant.

Den raad desheer Eijsschers Eijsch ende stucken daer nevens over gelegd, met aendagt gelesen en voorts op alles wel geleth hebbende, doende recht uijt de naem ende van

wegens de hoogmogende heeren staten generael der vereenigde nederlanden condemnneert de gevangne gebragt te werden ter plaetse, alwaer men gewoon is crimineele sententie te Executeeren en aldaer aen den scherpregher overgelevert zijnde agterwaerts op een kruijs gebonden, en van onderen op levendig gerabraekt tewerden voorts haar hooft met een bijl afgekapt, op een pen geset en zoo lange ten toon gestelt te werden tot dat de justitie omtrent alle de andere gevangens sal wesen g'Executeert om als dan het voors: hooft, benevens desselvs lichaem beneden het schavot met den puure verbrand te werden, ontsegt den heer Eijsscher sijnen anders genoomen Eijsch en condemnneert de gevangne daar en boven in de costen en uusen van justitie.

Aldus gearresteert ende gesententieert op dingsdag den 29,, november 1733./ bij den Edelen heer m,,r Cornelis van gaesbeeck tot thienhoven president / item de heeren m=r gerardus Gerlag m,,r Ioan Blaaukamer, m,,r Ioan rudolph Sappius, m,,r Jacob Lakeman, m,,r Bernard Iacob Dellafaille, m,,r Willem Vincent Helvetius/ george Fredrik Haderman, en Andries Herfort/ raeden van justitie mitsgaders gepronuntieert en g'Executeert ten daege ende ter presentie als in den hoofde deses.

Heden den 12,,e Decemb,,r 1733./ is voor het aen gaen van de Executie op de propositie van den heere president goet gevonden de assche van het lichaem van de gev: marthabientang naa dat het sal wesen verbrand naa dat het sal wesen verbrand naa het buijten geregt gebragt en aldaar onder de galsgraeven tewerden

Once the testing text is inputted into the trained model, Displycy is used to render the result. The rendered result is as follows:



Figure 9.7. The Rendered Result Using Displycy.

The result is satisfactory, with a few inaccuracies. For example, the word ‘*overgelevert*’ is labelled as PERSON, ‘*g’Executeert*’ as HONORIFIC, and ‘*Aldus gearresteert ende gesentiert*’ as ORG. We believe that a larger dataset can lead to more precise outcomes. The evaluation for this attempt is not included in this chapter due to the limited resources. However, the dataset and codebase for this experiment are publicly available on Kaggle.³³ The accuracy of the result can be seen in the following table:

Table 9.2. The accuracy of the NER results

Word	Automated Label	Accuracy	Correct Label
Martha Bintang	PERSON	Correct	
Wed:e	HONORIFIC	Correct	
Matthijs Lawalatoe	PERSON	Correct	
Gevange	PUNISHMENT	Correct	
Attentaat	FAC	Correct	
Moord	CRIME	Correct	
Ontfanger generael	OCCUPATION	Correct	
Ioachim Quilbant	PERSON	Correct	
Den Raad	ORG	Correct	
Eijsch	LAW	Correct	
Staten generael	ORG	Correct	
Verenigde Nederlanden	GPE	Correct	
Condemneert	LAW	Correct	
Scherpregter	OCCUPATION	Correct	
Overgelevert	PERSON	Incorrect	PUNISHMENT
Op een kruijs gebonden	PUNISHMENT	Correct	
Hooft met een bijl afgekapt	PUNISHMENT	Correct	
gevangen	LEGAL-STATUS	Correct	
Sal wesen	EVENT	Correct	
G’Executeert	HONORIFIC	Incorrect	LAW
Desselvs	PERSON	Correct	

³³ Asyrafi and Mahendra, ‘NER Dataset for VOC Archives Criminele Rollen’.

Word	Automated Label	Accuracy	Correct Label
Beneden het schavot met den puure verbrand te werden, ontsegt den heer	PUNISHMENT	Correct	
Eijsch	LAW	Correct	
Aldus gearresteert ende gesentieert	ORG	Incorrect	LAW
29 november	DATE	Correct	
1733	CARDINAL	Incorrect	DATE
M,,r	HONORIFIC	Correct	
Cornelis van geasbeecq	PERSON	Correct	
M=r	HONORIFIC	Correct	
M,,r	HONORIFIC	Correct	
Ioan Blaaukamer	PERSON	Correct	
M,,r	HONORIFIC	Correct	
Ioan Rudolph Sappius	PERSON	Correct	
M,,r	HONORIFIC	Correct	
Jacob Lakeman	PERSON	Correct	
M,,r	HONORIFIC	Correct	
Bernard Iacob Dellafaille	PERSON	Correct	
M,,r	HONORIFIC	Correct	
Willem Vincent Helvetius george	PERSON	Correct	
Fredrik Haderman	PERSON	Correct	
Andries Herfort/	PERSON	Correct	
Raeden van Justitie	ORG	Correct	
Gepronuntieert	PERSON	Incorrect	Not labelled
geExecuteert	ORG	Correct	
12,,e Decemb,,r	DATE	Correct	
1733	CARDINAL	Correct	
Den heere	PERSON	Correct	
Assche	LOC	Correct	
Gev:	LEGAL STATUS	Correct	

Word	Automated Label	Accuracy	Correct Label
Naa dat het sal wesen	PUNISHMENT	Correct	
Naa het buijten geregt gebragt	PUNISHMENT	Correct	

The model also missed some words that are supposed to be labelled. Based on the testing, the model has some difficulties recognising the custom entity ‘PUNISHMENT’. We suspect this is due to the length of the text describing the punishments. Additionally, the model missed the ‘LOCATION’ entity ‘*Amboina*,’ which is a standard spaCy label. This is likely because ‘*Amboina*’ is an old spelling for what is now known as Ambon. The name ‘*marthabientang*’ was also not recognised as a ‘PERSON,’ most likely due to a spelling error caused by HTR inaccuracy. Table 9.3 illustrates these issues:

Table 9.3. Unlabelled entities with their correct labels

Unlabelled Words	Correct Label
Amboina	LOCATION
Inlands Burger	LEGAL-STATUS
Gebragt te werden ter plaetse alwaer men gewoon is crimineele sententie te Executeeren	PUNISHMENT
Van onderen op levendig gerebraekt tewerden	PUNISHMENT
Op een pen geset en zoo lange ten toon gestelt te werden	PUNISHMENT
Marthabientang	PERSON

6. Conclusion

This study offers an innovative approach to reading and processing historical sources, specifically focusing on extracting historical data from early modern court records using supervised Named Entity Recognition (NER). Focusing on the *Criminele Rollen* from VOC-era Batavia, we demonstrated how NER, when carefully adapted, can yield meaningful structured data from unstructured colonial sources. Despite the specificity of our focus – capital punishment for enslaved persons in Batavia – the methodology developed here holds broader relevance. In doing so, this study offers both a) key insights into the challenges of historical data extraction and b) concrete recommendations for improving future digital humanities research.

One key insight from our research is that existing Handwritten Text Recognition (HTR) technology, such as the Transkribus HTR+ model used by the IJsberg project,

is not yet sufficient for immediate NER application. The structured yet inconsistent nature of *Criminele Rollen* entries – along with issues such as multilingual text, column formatting, and HTR errors – necessitates manual correction and supervision. While *Criminele Rollen* documents follow specific recording rules, any deviations can lead to disorganised HTR results, which hinder the effectiveness of the NER processes. The challenge we faced, in particular, was the extraction of the ‘PUNISHMENT’ label, which often appears in long and ambiguous forms. For example, the training corpus includes sentences describing multiple punishments, such as ‘...en van onderen op levendig gerabraekt tewaterden voorts haar hoofd met een bijl afgekapt,’ which translates to ‘...and from below, the living body is broken on the wheel before the head is beheaded with an axe.’ The machine struggles to label the correct length of the punishment phrase. In addition to that, it is hard for the machine to determine whether breaking on the wheel and beheading with an axe are one or two punishments in this verdict. This example illustrates the complexity of distinguishing between capital punishment and post-mortem punishment within a single phrase.

Despite facing several challenges, our research shows that a tailored NER pipeline for early modern Dutch texts is feasible and yields promising results. These challenges can be broadly divided into two categories: a) general challenges common to HTR-based corpora, such as multilingual text, irregular column formatting, and transcription errors, and b) task-specific challenges unique to this study, particularly the extraction of complex entity labels like ‘PUNISHMENT,’ which often appear in long, ambiguous sequences. This study successfully visualised the annotated data and tested our model using spaCy and Displacy, even if challenges remain in evaluating custom label performance due to gaps in spaCy’s documentation. These limitations point to the need for larger and more diverse training datasets, improved HTR outputs, and further experimentation with other NLP libraries that may offer better customisability.

More broadly, this project demonstrates how digital humanities tools can be used not only for technical exploration but also to recover voices long silenced in colonial archives. The structured extraction of personal names, legal statuses, and punishments allows researchers to trace patterns of marginalisation and legal violence embedded in bureaucratic language. While our case focuses on Batavia, the workflow is applicable to similarly structured archives elsewhere – be it in Southeast Asia, Africa, or the Americas – offering a reusable model for scholars working with early modern or colonial records. Future research should continue expanding the training corpus, improving model generalisation, and developing better tools for evaluating custom entity recognition. Interdisciplinary collaboration between historians, linguists, and data scientists will be essential for unlocking the full potential of machine learning in historical research.

