Exploring text mining techniques to structure a digitised catalogue

August 2019

Karen Goes
2624586
k.w.m.goes@student.vu.nl

VU University Amsterdam
Faculty of Humanities
Research Master Humanities
Human Language Technology

First supervisor: Dr. E. Maks
Second supervisor: Sara Veldhoen MSc
Second reader: Dr. H. D. van der Vliet

Contents

Abstract

This research aims to obtain structured data from digitised Brinkman catalogue volumes and identifying which text mining techniques can be used for this task. The Brinkman catalogue lists the books, journal titles and maps published in the Netherlands since 1846. They have been digitised by the National Library of the Netherlands using Optical Character Recognition (OCR). However, this data is unstructured, uncorrected and cannot be processed computationally. Since the data is uncorrected an analysis is performed to identify which volumes meet the requirements to be used for further processing. The entries are then formed by identifying the start and end of the entry. Each entry contains metadata about a book such as the author, title, publisher and retail price. This metadata is extracted from the entries using different text mining techniques namely, a rule-based system including regular expressions, a probabilistic context-free grammar, and named entity recognition. The extracted information is improved using external knowledge such as a list of Dutch surnames for authors and Dutch and Belgian city names. For the three most recent volumes the National Library has provided evaluation data that is already digitally available. To evaluate the remaining volumes evaluation data has been manually created for 100 entries. From the evaluation results the conclusion is drawn that the rule-based method with regular expressions is the best technique. The extracted data is transformed to a standard format for bibliographic metadata, which makes it structured and searchable data.

1 Introduction

Over the past two decades a ton of books have been published in and about the Netherlands. The Brinkman's Catalogue of Books seeks to list all these books, divided over multiple volumes. The volumes give an overview of all these books published between 1833 and the present day. They are available as printed books (up to 2002) and later on as CD-ROMs (starting in 2002) (Veen & Waterschoot, 2001, p. 10).

For this research project the volumes covering the years 1833-1980 are processed. The volumes have previously been scanned and ran through an Optical Character Recognition (OCR) process, commissioned by the National Library of the Netherlands. The outputs from the OCR process form the base of this research. The OCR output is one big text file with no structure in it, which means that it is not in a machine friendly format. The goal of this research is to transform the OCR output to a structured machine friendly format.

1.1 Task

The task at hand involves processing the digitalised Brinkman catalogue volumes, extracting information from them and converting this to a structured PICA+ format. The PICA+ format is a format in the catalogue system that is used at the National Library, the 'Gemeenschappelijk Geautomatiseerde Catalogiseersysteem' (GGC).

To accomplish this the bibliographical entries from the volumes need to be recreated and their metadata should be extracted using automatic processing. The extracted data is then transformed to the PICA+ format.

The goal and task of this research is to create and evaluate a method to extract specific bibliographical meta data from the catalogue volumes.

1.2 Research question

The research question that is formulated to guide the research reads:

Which text mining techniques can be used to structure digitised bibliographical data and what is the best way to evaluate these methods?

To answer the first part of the research question multiple text mining techniques will be used to extract bibliographical metadata, such as *author*, *title* and *publisher*, from the OCR output. Three techniques will be used: Splitting & Regular Expressions, Named Entity Recognition and a Probabilistic Context-Free Grammar. These techniques will all be evaluated and compared, to find out which method is the most efficient and accurate for this task.

For the second part of the research questions two different evaluation sets will be tested. Part of the data set will be evaluated using manually created data sets based on the original Brinkman Catalogue volumes. The remaining volumes in the data set will be evaluated using already digitally available data from the library catalogue of the National Library.

1.3 Motivation

The research topic for this thesis has been suggested by the National Library of the Netherlands, where I did an internship to conduct the research. This topic is of interest to the National Library because the data from the Brinkman catalogues is currently only available as PDF scans and uncorrected OCR output. This means that it is not machine readable and not available as structured data.

The Brinkman catalogues, as mentioned before, contain all the books published in and about the Netherlands from 1833 till the present. The National Library would like to have a copy of all of the books mentioned in these catalogues. However, the library has only been actively collecting all the books that have been published in and about the Netherlands since 1974. Consequently, books are missing from their collection from before that date.

Because the Brinkman catalogues give an overview of all the books that have ever been published in the Netherlands, this data is valuable to the National Library. With this data they can compare their current collection with the list of books from the Brinkman catalogues to identify which books are missing from their collection. To be able to do this, the data from the catalogues has to be available in a machine readable and structured format.

If the method proves successful, it could be used for other similar data, such as auction catalogues.

2 Related work & Background

This section will give an overview of the work that has been done related to the current research as well as some background on the used techniques. No literature was found about bibliographical catalogues, therefore the work related to another structured material will be discussed, namely a dictionary. The entry of a dictionary is similar to that of a bibliographical entry in the catalogue since both have a consistent internal structure with the lexeme, part of speech, definitions and possibly translations. Due to this similarity the approaches used to tackle problems such as typographic errors, finding the beginning and the end of the separate entries and extracting information from them are of interest for the current research. The methodologies used in these research projects can be used for the current research, may it be in an adapted form.

The works that will be discussed below in more detail are: Maxwell and Bills (2017), Khemakhem et al. (2017), Ma et al. (2003), Bago and Ljubešić (2015) and Karagol-Ayan et al. (2003).

2.1 Digitising and structuring dictionaries

A case study involving the digitisation of printed dictionaries of endangered languages is described in a paper by Maxwell and Bills (2017). Three problems are mentioned that occur when a printed dictionary is digitised (Maxwell & Bills, 2017, p. 85).

1. *Typographic errors*: Errors in the data caused by the inability of the OCR software to recognise certain characters, structures or entire paragraphs.
2. *Conversion to a structured machine-readable format*: How is the printed dictionary processed to be transformed into a machine-readable format. This includes the process of obtaining the lexical entries within the dictionary.
3. *Extract details from lexical entries*: How are the details from the lexical entries extracted and transformed into a structured (standard) format.

The problems are also met by the current research into structuring a digitised catalogue. Therefore, I discuss below how multiple research projects on digitising dictionaries deal with these problems and the evaluation, also for the current research, are discussed below.

2.1.1 Typographic errors

Typographic errors, or OCR errors, are errors made created by the OCR software when certain characters or even entire columns are not recognised correctly. This can cause incomplete and unintelligible output, which then influences the results of the research. Maxwell and Bills (2017) do not try to fix the typographic errors, they consider this as a nearly last step, rather than an initial step. Since it is unlikely to find an off-the-shelf OCR model that can deal with rare letters that you come across in a minority language. Therefore, no OCR errors are manually fixed.

Another research that aims to structure a digitised dictionary is by Khemakhem et al. (2017). In their research, they run their models on two different dictionaries: a digital dictionary and a digitised dictionary. Since the first dictionary is digitally born, no OCR processing is needed and therefore there are no typographical errors. The latter dictionary is run through OCR software, they mention that their OCR quality is of relatively good quality but still presents some anomalies. It is not described how these anomalies are dealt with.

Ma et al. (2003) work on parsing and tagging bilingual dictionaries. They make use of trainable techniques to recognise special symbols and to correct symbols the OCR fails on. No specific trainable technique is mentioned for this problem, however, Gabor filters are mentioned as the technique used to recognise fonts. It can be assumed that the same technique is implied for the correction of OCR errors.

The 18th century dictionary used in the research by Bago and Ljubešić (2015) is process with OCR software. The OCR errors are manually corrected by undergraduate students, by checking them against the corresponding pictures.

The OCRed bilingual dictionaries in the research by Karagol-Ayan et al. (2003) on machine translation lexicons are constructed by three different methods. One of these methods, the rule-based one, accounts for OCR errors by allowing for a relaxed matching of OCRed output to information types.

A similar approach to that of Maxwell and Bills (2017) is chosen for the current research where the focus is not to fix OCR errors but rather work with them and around them. Still some OCR errors are fixed both manually and automatically. This concerns common errors such as the letter 'L' being replaced with a '1'.

Research by Traub et al. (2015) shows that OCR errors can have an im-

pact on the research results. They interviewed scholars asking them how the OCR quality influenced their research and conducted a literature study on the impact of OCR quality on scholarly research. The interviews revealed that the uncertainty about the quality of the OCR was seen as a serious obstacle to wider the adaption of digital methods in the humanities. Furthermore, most scholars were unable to quantify the impact of OCR errors on their own research tasks. The conclusion drawn from the literature study is that even though OCR quality is a widely studied topic, it is mostly on the performance of an OCR tool. Their analysis shows that for most research tasks the problem cannot be solved with better but still imperfect OCR software.

2.1.2 Conversion to a structured machine-readable format

The papers mentioned above use different techniques to convert the often printed dictionaries to a structured machine-readable format and to identify the lexical entries. The most popular technique used to get a digital version of the printed dictionary is with the use of OCR software. It is used by Maxwell and Bills (2017), Khemakhem et al. (2017), Ma et al. (2003), Bago and Ljubešić (2015), Karagol-Ayan et al. (2003), and also in the current research. This results in a text file with plain text, the content of the printed dictionary.

The research by Maxwell and Bills (2017) uses a system that has a different type of OCR output, a hOCR output (Breuel, 2007) format. This is in an XML format and already gives information about hypothesised paragraphs, columns, words, and lines, and additional information about the font and font style. However, this is not enough to get correctly formed lexical entries. To accomplish this post-processing of the hOCR output is necessary in the form of a Python script in which the user supplies several parameters. These parameters include, among others, the number of columns on a page, any information that spans columns such as headers and/or footers, the level of indentation in a lexical entry and an approximate measure of each indent. After running this script, the output is an XML file whose structure below the root element consists of a sequence of inferred lexical entries.

In their research Ma et al. (2003) use a three-step method, applied to the OCR output, to do the lexical entry segmentation. The result is a set of lexical entries that represent a physical partition of a page from the dictionary.

1. *Feature Extraction and Analysis*: Extraction of physical characteristics

which indicate an entry boundary.
2. *Training and Segmentation*: Learning the parameters of an entry segmentation model.
3. *Correction and Bootstrapping*: Feedback from an operator, who makes corrections to a small subset of the results that contain errors. Using the corrected segmentation results, bootstrapping samples are generated and used to retrain the system.

This system from Ma et al. (2003) is used in the research by Karagol-Ayan et al. (2003) to identify the lexical entries. For the purpose of their research it is assumed that the output of the system contains the following characteristics:

- Each page is divided into dictionary entries
- Each entry is associated with an entry type
- For each entry, lines and tokens are identified
- For each token, font style is provided

Khemakhem et al. (2017) use a different approach to identify the lexical entries in their digital and OCRed digitised dictionaries. They use a supervised machine learning system called GROBID (GeneRation Of Bibliographic Data) (Romary & Lopez, 2015), which is used for parsing and extracting bibliographic metadata from scholar articles and adapted for the specificity of the use case of digitised dictionaries. The Conditional Random Fields (CRF) algorithm performs a multi-level sequence labelling of text blocks in a cascade fashion which are then extracted and encoded in TEI (Text Encoding Initiative)[1] elements, where the various segmentation levels are associated with an appropriate XML level.

The approach for the current research is partially overlapping with that of Bago and Ljubešić (2015), where the beginning and end of each lexical entry is manually annotated.

The current research involves manual annotation and automatic identification of bibliographical entries using the alphabetical nature of the catalogue. This differs from the other research projects mentioned, none of these projects use the alphabetic properties of the dictionary. The approaches described could not be applied for the current research due to the lack of rich OCR output, also the different volumes contain different structures for the

[1] https://tei-c.org/

entries which would make it difficult to create a machine learning algorithm that can identify the start and end of an entry.

2.1.3 Extract details from lexical entries

The result of creating a machine readable digitised dictionary is a list of lexical entries. The next step is to extract the details from them and save this in a machine-readable (standard) format. The main techniques used for this step are using a machine learning system, a grammar and a rule-based system.

The machine learning techniques to extract details from the lexical entries are used by Khemakhem et al. (2017), Bago and Ljubešić (2015), Ma et al. (2003) and Karagol-Ayan et al. (2003).

Khemakhem et al. (2017) and Bago and Ljubešić (2015) use the CRF sequence labelling algorithm. Like the identification of the lexical entries, Khemakhem et al. (2017) uses the GROBID system, on a deeper level, to extract information from the lexical entries. It uses different models for the form, senses, and grammatical group of a lexical entry. The results are transformed into an XML format with TEI elements.

Bago and Ljubešić (2015) describe different features that are used by the CRF machine learning algorithm, including:

- *token:* a token in its original form
- *ltoken*: lowercased token
- *lcasebool*: a Boolean variable whether a token is lowercased or not
- *prevNtoken* and *nextNtoken*: N tokens before and after a specific token, for N = 1..4
- *prevNlcasebool* and *nextNlcasebool*: a Boolean variable whether N tokens before and after are lowercased
- *lang*: a language label of the token
- *suffixN*: a suffix of a specific token of length N=4

The format of the resulting output is not mentioned in the paper.

Ma et al. (2003), and Karagol-Ayan et al. (2003) use the Hidden Markov Model (HMM) machine learning algorithm as one of their methods, besides a rule-based one. This algorithm works with hidden states, observable states and different probilities (Manning & Schütze, 1999).

In the paper by Ma et al. (2003) the observable states are the tokens and the hidden states the tags corresponding to the tokens. The algorithm uses six features which are mentioned below.

- *Content*: Category of the keyword if the token is a keyword, SYM if it is a special symbol, NUM if it is a number, and 'null' otherwise.
- *Font*: Font style of the token (normal, bold, italic, etc.).
- *Starting symbol*: Special punctuation mark if the token begins with one, 'null' otherwise.
- *Ending symbol*: Special punctuation mark if the last character of the token is one, 'null' otherwise.
- *Second ending symbol*: Special punctuation mark if the second to last character of the token is one, 'null' otherwise.
- *Is-first token*: True if this is the first token of an entry, false otherwise.

The rule-based system uses the font style and separators as dividers, separators include a semicolon, brackets, full stops and commas. The tokens in an entry are grouped into elements that will be tagged as a unit. The output of the systems is converted to a HTML format, which preserves the nested representation present in a dictionary.

Karagol-Ayan et al. (2003) use the same features and approach as Ma et al. (2003), with the addition of one feature: Is-Latin content, which describes whether or not the content consists of Latin based characters. The same holds for the rule-based method where no adaptions are made. No specific output format for the final lexicon is mentioned.

The research by Maxwell and Bills (2017) uses a final state grammar with regular expressions (Thompson, 1968). This grammar uses rules to parse the entry based on the tokens, for example, a lexical entry consists of a headword, a part-of-speech tag and one or multiple gloss words. The resulting structured lexical entries are transformed to an XML format.

The current research uses a similar approach, where the regular expressions are used in a rule-based method and a different type of grammar is used, a probabilistic context-free grammar. This type of grammar has rules similar to the final state grammar, with the addition of a probability to each rule. The probability feature of this grammar can be deployed to deal with some of the OCR errors that are present in the OCR output of the catalogue volumes. The rule-based approach uses separators, like the other rule-based systems, and regular expressions to extract consistent elements. Due to the

lack of annotated data it was opted not to use a machine learning approach for this research. The resulting extracted information is transformed into a PICA+ format.

2.1.3.1 Named Entity Recognition (NER)

Another method that is used in the current research is Named Entity Recognition. This method extracts Named Entities from text based on features such as capitalisation, previous and next tokens and type of token. After they are recognised they are classified into different categories such as person, location or organisation, this is called Named Entity Recognition and Classification (NERC). The study by Nadeau and Sekine (2007) gives an overview of the literature about the field of NERC, from 1991 to 2006. The recognition and classification of named entities in the current research is done by the Dutch language model in the natural language processing tool SpaCy [2].

2.1.4 Evaluation

The evaluation methods for the research projects are similar across most of them, namely by manually annotation a number of pages of the dictionary. The research by Ma et al. (2003) manually annotates 5 pages for their evaluation. Karagol-Ayan et al. (2003) also manually annotates 5 pages as one of their evaluation methods, for another method they manually annotate 100 random entries from the dictionary. Their evaluation results are categorised in three categories:

1. *Missing*: Not in the generated lexicon
2. *Extra*: Not in the original dictionary
3. *Incorrect*: Tagged correctly, but incorrect because of OCR noise

The same evaluation method is used for the current research where 100 entries from randomly chosen pages are manually annotated and categorised into the three categories, per piece of information.

Khemakhem et al. (2017) have also chosen the approach of manually annotating a certain amount of pages, varying per dictionary and segmentation level, to evaluate their system's performance.

Bago and Ljubešić (2015) mention using a 10-fold cross-validation method to evaluate the performance of each parameter.

[2] https://spacy.io/models/nl

2.2 History of the Brinkman catalogues

The Brinkman catalogue of books first originated in 1846, when Carel Leonard Brinkman (1820-1881) first put out and advertisement in a newspaper for book stores. In this advertisement he called upon publisher in the Netherlands to send him details about books and dissertations that were not officially published, to be recorded in a list he called the 'Alphabetische naamlijst van boeken' (Veen & Waterschoot, 2001, p. 15).

The idea of creating such a list was not new, since the Dutch bookseller A. B. Saakes started publishing a monthly list from 1790 called 'Naamlijst van Nederduitsche boeken'. The publication of this list was discontinued in 1811 under the French rule. Later on other lists were published for 1814-1819 and 1820-1824. In 1832 the brothers Van Cleef attempted to summarise the messy period in a list called the 'Alphabetische naamlijst van boeken' which was based on the list published bij Saakes.

With the first list of books published by Brinkman in 1846, he tried to close the gap by publishing the 'Alphabetische naamlijst van boeken, plaat- en kaartwerken' in 1858, which covered the years 1833-1849 (Veen & Waterschoot, 2001, p. 15-16).

The next large catalogue that was published that spanned multiple years was the 'Brinkman's Catalogus der Boeken, Plaat- en Kaartwerken die gedurende de jaren 1850-1882 in Nederland zijn uitgegeven of herdrukt'. Brinkman's son-in-law Rimmer Reinders van der Meulen took over the business together with Brinkman's son in 1878. This did not last very long and in 1885 another firm took over. They have the 'Naamlijst' to another publisher that would publish the list for nearly a century: A. W. Sijthof (Veen & Waterschoot, 2001, p. 21).

In the 1960s the work on the Brinkman catalogues became too much for the Brinkman department of the publisher. A solution presented itself in the form of the National Libary of the Netherlands. In 1974, the 'Depot van Nederlandse Publicaties' was created on an experimental basis, to collect all the publications that needed to be included in the catalogue, as well as for the library itself. The Brinkman catalogue was expanded to include a bibliography of (semi-)governmental publications, which had already been published by the library since 1929. In 1982 the National Library of the Netherlands took over the publishing rights and responsibility for the Brinkman catalogues (Veen & Waterschoot, 2001, p. 32-35).

3 Data

The data consists of different volumes of the Brinkman Catalogue of Books and is retrieved from the Digitale Bibliotheek voor de Nederlandse Letteren (DBNL) website.[3] The volumes are available as digitised PDF scans and text files (by means of OCR) of printed books, and have originally been published between 1833 and 1980. A total of 31 volumes are retrieved to be used in this research.

A collection of volumes of the Brinkman Catalogue of Books is also digitally available on the Delpher website.[4] However, this only concerns year books opposed to the volumes containing multiple years on DBNL. The volumes from DBNL contain more information and most importantly this information has the same structure across the volume. The year books on Delpher only contain one year each, while the majority of the volumes from DBNL consist of multiple years. When these bigger volumes are used, the structure and layout is the same for the different years in this volume, unlike different layouts and structures across the year books. Therefore using the volumes from DBNL makes the process of extracting bibliographical metadata easier and more efficient because the same method can be used on multiple years.

A volume contains bibliographical information on the books published in one or multiple years. The earlier volumes are larger and cover more than one year. The largest volume includes books published between 1850 and 1882. Volumes in the range 1911-1965 cover five years each, whereas the most recent volumes only cover a single year. Also, some volumes are divided into two parts based on either the alphabet (A-M, L-Z) or time period (January-June, July-December).

Most volumes consist of three sections: a catalogue of books, a title catalogue and a topic register. The sections contain the same information as the bibliographical list of publications, but they are ordered either by title or topic, as opposed to author (if the author is known). For the purpose of this research, these additional sections are not processed and only the catalogue of books is used.

Due to the different periods of time in which the volumes were originally published, there is a big variance in the format, layout and content across the volumes. The content of each volume consists of a list of bibliographical

[3] https://www.dbnl.org/auteurs/auteur.php?id=brin003
[4] www.delpher.nl

entries and detailed information about them, but the amount of information given varies across different volumes. For example, not all volumes give the International Standard Book Number (ISBN) of a bibliographical entry, for which the reason is that ISBN was not standardised till 1970.[5] Likewise, the *number_of_pages* of a bibliographical entry is not included in every volume.

Nonetheless, the *title*, *author*, *publisher* and *city_of_publication* are found among all the volumes. Another similarity across all the volumes is that each entry starts with the name of the author and entries are ordered alphabetically. If the author is unknown, the entry starts with the title of the book. An overview of what bibliographical metadata can be found in which volume, for the volumes selected in Section 4.1, is shown in Table 1.

The information given in the table concerns the bibliographical metadata, described below. These 'tags' are used throughout the thesis when discussing the metadata.

- *Author*: The author of the bibliographical entry
- *Title*: The title of the bibliographical entry
- *City_of_publication*: The city in which the bibliographical entry is published
- *Publisher*: The publisher that published the bibliographical entry
- *Year_of_publication*: The year in which the bibliographical entry is published
- *Size_of_book*: The size of the bibliographical entry
- *Price_of_book*: The retail price of the bibliographical entry
- *Number_of_pages*: The number of pages that the bibliographical entry is made up of
- *ISBN_number*: The ISBN number of the bibliographical entry
- *Production_number*: The Brinkman production number of the bibliographical entry

Other relevant information about the data concerns the use of dashes for repeated metadata, and the placement of bibliographical entries that start with a 'IJ' in the alphabetical order of the volumes. Dashes used to indicate repeated metadata are used up to 1975. The bibliographical entries that start with a 'IJ' are placed at the 'Y' in the alphabetical order for all volumes, except for the first two (1850-1882, 1882-1891).

[5] https://www.isbn.org/ISBN_history

Volume		Bibliographical metadata in catalogue				
		Author	*Title*	*City_of_publication*	*Publisher*	*Year_of_publication*
Group 1	1850-1882	X	X	X	X	X
	1882-1891	X	X	X	X	X
	1931-1935	X	X	X	X	X
Group 2	1936-1940	X	X	X	X	X
	1961-1965	X	X	X	X	X
	1971	X	X	X	X	X
	1975	X	X	X	X	X
Group 3	1979	X	X	X	X	X
	1980	X	X	X	X	X

Volume		Bibliographical metadata in catalogue				
		Size_of_book	*Price_of_book*	*Number_of_pages*	*ISBN_number*	*Production_number*
Group 1	1850-1882	X	X			
	1882-1891	X	X			
	1931-1935	X	X	X		
Group 2	1936-1940	X	X	X		
	1961-1965	X	X	X		
	1971	X	X	X	X	X
	1975	X	X	X	X	X
Group 3	1979	X	X	X	X	X
	1980	X	X	X	X	X

Table 1: Overview of which bibliographical metadata is present in which volume

Within the scope of this research, we start from the text files with the OCR output. The PDF scans are used for consultation in the processing phase and evaluation, described in sections 4.3 and 5, respectively.

4 Method

This section describes the main research method that consists of a number of steps, beginning with selecting the appropriate volumes which can be processed automatically and ending with a file with structured PICA+ data of the book entries for each volume. To accomplish this the quality of the OCR output of each volume is assessed for which the book entries will be created. Next, the bibliographical metadata from these book entries is extracted and enriched with external knowledge. As a last step the extracted data is transformed to a structured PICA+ format. In Appendix B, a list of the deliverables of this research is given as well as where they can be found. An overview of this method can be seen in Figure 1.

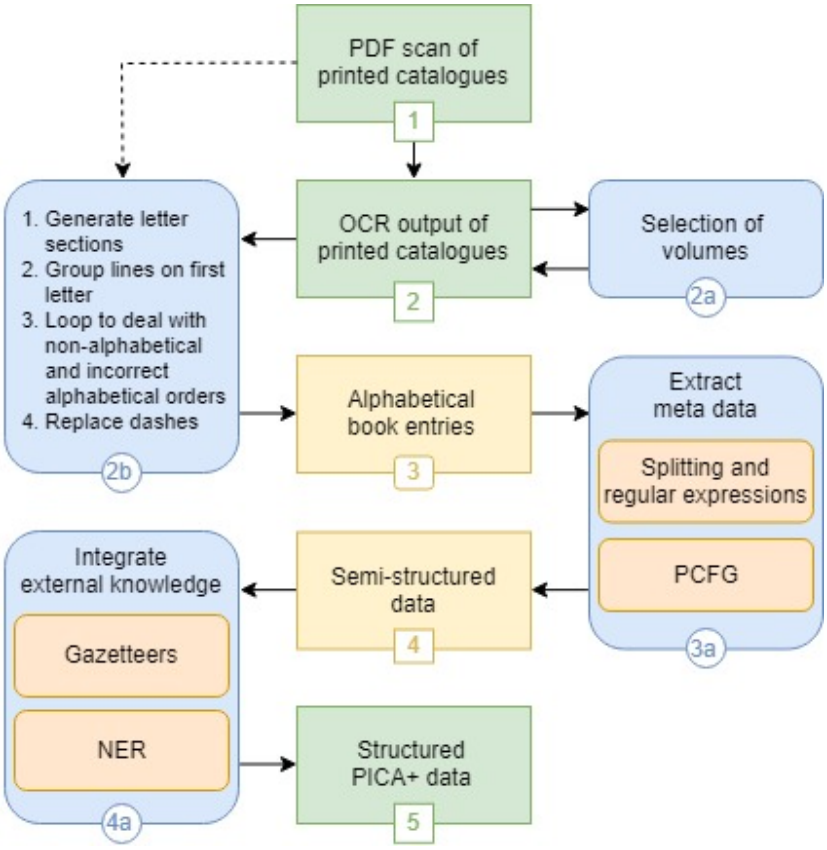

Figure 1: Schematics of the main steps of the method used to transform the OCR output of catalogue volumes to structured PICA+ data.

Bibliographical entries that refer to other entries are disregarded. An example of such a reference can be seen below in (1), in which the entry of the *author* refers to the full bibliographical entry, where the complete information concerning the entry can be found.

(1) Aalders, C., Zie Luisterend leven.

The OCR process is not faultless and the output contains errors that may effect my research. Minor errors concern for example misspellings of a word. However, major errors include the incorrect recognition of the layout of a page causing incomplete entries. Because of these issues it is important to have a close and critical look at the output of the OCR before defining further steps. This qualitative analysis is the first step in my research method and is described below in section 4.1.

4.1 Selection of volumes

To assess the quality of the OCR output for the different catalogue volumes, a qualitative analysis is performed. The goal of this analysis is to identify which volumes can be processed automatically to be able to extract the metadata from the book entries.

When looking at the output, the following three issues need to be taken into account:

1. The layout of the catalogue should be correctly recognised by the OCR process and written to the OCR output, otherwise, it will produce incomplete bibliographical entries. This mistake can be seen in Figure 2 where the two columns present in the original catalogue are not processed correctly in the OCR output. In order to automatically process the OCR output line by line the columns should be below each other, rather than next to each other. If the columns are next to each other one line contains two partial and different bibliographical entries. When forming the bibliographical entries this will result in incomplete entries that combine two originally different entries. In the extraction process this would lead to incorrect bibliographical metadata.

2. The OCR of the content of the catalogue volumes has to be inspected. In case of a lot of spelling errors and/or the frequent occurrence of symbols in the bibliographical content, the volumes have to be disregarded.

Group 1	Group 2	Group 3
1979	1931-1935	1850-1882
1980	1961-1965	1882-1891
	1971	
	1975	

Table 2: Overview of the three created groups and which volumes belong to which group

If these misspellings and symbols were to be used, the extracted information would possibly be illegible and not usable for the purpose that the National Library has for the extracted metadata. This is shown in Figure 3.

3. In older volumes, up to 1975, dashes are used as a replacement for repeating information such as the *author* or the first word of a *title*. These dashes should be visible in the OCR output to be able to extract the correct *author* or *title*. An example of this issue can be seen in Figure 3, in which the scan of the original catalogue shows dashes where the *author* is repeated. However, these dashes are not transferred to the OCR output. Without the dashes in the OCR output it is not possible to extract the correct *author* or *title* of a bibliographical entry that starts with a dash, because there are no other indications that information is missing.

After the qualitative analysis, nine volumes are selected for further processing. These volumes meet the requirements with respect to the quality of the OCR output. It concerns the following volumes: 1850-1882, 1882-1891, 1931-1935, 1936-1940, 1961-1965, 1971, 1975, 1979 and 1980. The last volume is not complete, the OCR output stops at the beginning of the letter P. However, the quality of the output meets the remaining requirement which means that the catalogue can be used even though it is incomplete.

These volumes have been grouped based on structure of the bibliographical entries. The order in which the metadata is written within the bibliographical entries differs, also the punctuation that separates the types of metadata are not the same across the volumes. When this is equal for volumes within one group they can be processed similarly. The *year_of_publication* is also a factor in the grouping process. See Table 2 for the created groups.

Aa (M. W. van der), Zie Meerbeke.

Aafjes (J. D.),. Buig de stem ! Leesboek ter bevordering van den goeden leestoon. Amst., W. Versluys. 1912. post 8o. ƒ 0.30.

—— Practisch zaakonderwijs. Wordings-rijen in natura van in het dagelijksch leven veelvuldig voorkomende artikelen. Zutphen, W. J. Thieme & Co. 1913.

1e rij. Glas (blazen, slijpen en etsen). 20 onderdeelen. ƒ 6.—.

2e rij. Borstelwerk (het maken van een stoffer). 7 onderdeelen. ƒ 3.—.

3e rij. Goudsche pijpen. 8 onderdeelen ƒ 3.—.

4e rij. Lucifers. 20 onderdeelen. ƒ 3.—.

5e rij. Muurtegels. 8 onderdeelen. ƒ 4.—.

6e rij. Cacao. 7 onderdeelen. ƒ 4.—.

Handleiding ƒ 0.30.

—— Veel plezier!! Leesboek voor de lagere school. Amst., W. Versluys. 1912. 2 stukjes. post 8o. elk stukje ƒ 0.25.

Aalderen (B. W. van), Nieuwe wijsjes. Twaalf tweestemmige liedjes voor de lagere school. Gron., P. Noordhoff. 1911. gr. 8o. ƒ 0.35.

Aalderink (H.), De zoetwatervisschen in Nederland en de kunst om ze te vangen. 2e, geheel herziene druk. Met [8] platen. Rott., D. Bolle. 1911. gr. 8o. m. 3 tab. ƒ 2.25; geb. ƒ 2.75.

Aalders (G. C.), De valsche profetie in Israël. (Proefschrift, vrije univ. Amsterdam). Wageningen, J. Zomer, Drukkerij „Vada". 1911. gr. 8o. ƒ 2.50.

—— Sporen van animisme in het Oude Testament? Kampen, J. H. Kok. 1914. gr. 8o. ƒ 0.60.

Aalders (W. J.), Christendom en cultuur. 1911. Zie Levensvragen. Ve Serie, No. 4.

—— De waarde der mystiek voor ons geloofsleven. 1914. Zie Levensvragen. VIIe serie. No. 7.

(a) A catalogue with a two column layout

```
Aa (M. W. van der), Zie Meerbeke.

/lanes (J. D.),. Buig de stem I Leesboek
ter bevordering van den goeden leestoon.
Amst., W. Versluys. 1912. post 80. / 0.30.

-       Practisch zaakonderwijs. Wor-
dings-rijen in natura van in het dagelijksch
leven veelvuldig voorkomende artikelen.
Zutphen, W. J. Thieme & Co. 1913.
       le rij. Glas (blazon, slijpen en etsen).

    20 onderdeelen. f 6.-.

       2e rij. Borstelwerk (het maken van

    een stoffer). 7 onderdeelen. f 3.-.

       30 rij. Goudsche pijpen. 8 onder-
    deelen f 3.-.
       4e rij. Lucifers. 20 onderdeelen. 13.-.
       5e rij. Muurtegels. 8 onderdeelen.
      4.-.
       6e rij. Cacao. 7 onderdeelen. f 4.-.
    Handleiding f 0.30.
-      Veel plezier I I Leesboek voor de
lagere school. Amst., W. Versluys. 1912.
2 stukjes. post 80. elk stukje 1 0.25.

   Aalderen (B. W. van), Nieuwe w(jsjes.

Twaqlf tweestemmige liedjes voor de lagere
school. Gron., P. Noordhoff. 1911. gr. 80.
f 0.35.

   Aalderink (H.), De zoetwatervisschen in
Nederland en de kunst om ze te vangen.
2e, geheel herziene druk. Met [8] platen.
Rott., D. Bolle. 1911. gr. 80. m. 3 tab.
f 2.25 ; gob. f 2.75.

   Aalders (G. C.), De valsche profetie in

Israel. (Proefschrift, vrije univ. Amsterdam).

Wageningen, J. Zomer, Drukkerij ,,Vada".

1911. gr. 80. f 2.50.

     Sporen van animisme in het Oude
Testament ? Kampen, J. H. Kok. 1914.
gr. 80. f 0.60.

   Aalders (W. J.), Christendom en cultuur.
1911. Zie Levensvragen. Ve Serie, No. 4.
   -      De waarde der mystiek voor ons
geloofsleven. 1914. Zie Levensvragen. Vile
serie. No. 7.
```

(b) OCR output of the catalogue snippet in (a)

Figure 2: A catalogue with a two column layout which is not correctly recognised by the OCR process

Aa (A. J. van der), Aardrijkskundig Woordenboek van Noord-Braband. *Breda, van Gulick en Hermans.* 1833. 12°. met kaartje. . . . . *f* 2,60.

——— Herinneringen uit het gebied der geschiedenis betrekkelijk de Nederlanden. *Amst. J. C. van Kesteren.* 1835. gr. 8°. . . . . *f* 3,60.

——— Nieuwe herinneringen uit het gebied der geschiedenis betrekkelijk de Nederlanden. *Amst. J. C. van Kesteren.* 1837. gr. 8°. . . . . *f* 3,60.

——— Lees- en vertaalboekje voor de hoogste klassen der Fransche scholen, met eene woordenlijst, waarin de vertaling der daarin voorkomende woorden opgegeven wordt. *Amst. Schalekamp, van de Grampel en Bakker.* 1836. kl. 8°. . . . *f* 0,60.

——— Zamenspraken in de Nederd. en Fransche talen, met woordelijke overzetting van elken zin. *Breda, van Gulick en Hermans.* 1836. kl. 8°. . . . . *f* 0,25.

——— Geschied- en aardrijkskundige beschrijving van het koningrijk der Nederlanden en het Groot-hertogdom Luxemburg. *Gorinch. J. Noorduyn en Zn.* 1841. gr 8° *f* 4,80.

——— Beknopte geschied- en aardrijkskundige beschrijving van het koningrijk der Nederl. en het Groot-hertogdom Luxemburg; benevens een kort overzigt van Nederl. bezittingen buiten Europa. *Gorinch. J. Noorduyn en Zn.* 1844. 3e verm. dr. 1848. kl. 8°. *f* 0,17½.

——— China en zijne bewoners geschetst voor jonge lieden. *Amst. G. J. A. Beijerinck.* 1845. kl. 8°. met houtsneêpl. . . . . *f* 1,80.

——— Geschiedkundige beschrijving van de stad Breda en omstreken. *Gorinchem, J. Noorduyn en Zoon.* 1845. gr. 8°. . . . . *f* 3,90.

——— Nederlands Oost-Indië, of beschrijving der Nederl. bezittingen in Oost-Indië, voorafgegaan van een beknopt overzigt van de vestiging en uitbreiding der magt van Nederl. aldaar. *Amst. J. F. Schleijer.* 1845—49. roy. 8°. met pl. en kaarten. 1e—19e afl. Elke afl. *f* 0,50.

(a) A snippet from the 1833-1849 catalogue where the repeated occurence of the same *author* is indicated with a dash

```
♠.111u (A. J. VAN DER), Aardrijkskundig Woordenboek van Noord-Braband. Breda, van
Gulick en Hermans. 1833. 12°. met kaartje.      •       f 2,60.
Herinneringen uit het gebied der geschiedenis betrekkelijk de • Nederlanden. 4mst.
J. C. van Ifesteren. 1835. gr. 8°.      f 3,60.
Nieuwe herinneringen uit het gebied der gesc• hieden• is be• trekkelijk • de Nederlanden.
.Amst. J. C. van I-esteren. 1837. gr. 8°.       f 3,60.
Lees- en vertaalboekje voor de hoogste klassen der.Fransche seholen, met eene
woordenlijst, waarin de vertaling der daarin voorkomende woorden opgegeven wordt.
,dmst. Schalekamp , van de Grampel en fakker. 1836. kl. 8°. .   •       f 0,60.
Zamenspraken in de Nederd. en Fransche talen, met woordelijke overzetting van
elken zin. Breda, van Gulick en Germans. 1836. kI. 8°. .        f 0,25.
Geschied- en aardrijkskundige beschrijving van het koningrijk der • Nederlanden en
het Groot-hertogdom Luxemburg. Gorinch. J. Noorduyn en Zn. 1841. gr 8° f 4.80.
Beknopte geschied- en aardrijkskundige beschrijving van het koningrijk der Nederl.
en het Groot-hertogdom Luxemburg; benevens een kort overzigt van Nederl. bezittingen
buiten Europa. Gorinch. J. Noorduyn en Zn. 1844. 3e verin. dr. 1848. kl.8°. f0,17I.
China en zijne bewoners geschetst voor jonge lieden. Amst. G. J. A. Beijerinck.
1845. kl. 8". met houtsneepl.   f 1,80.
      Geschiedkundige beschrijving • van • de sta• ll Breda • en oms• treken. • Gorinchem, J.
foorduyn en Zoon. 1845. gr. 8°. f 3,90.
Nederlands Oost-Indie, of beschrijving der Nederl.~bezittingen in Oos• t-Indie, voor-
afgegaan van een beknopt overzigt van de vestiging en uitbreiding der magt van Nederl. aldaar.
dmst. J. F. Schle~er. 1845-49. roy. 8°. met pl. en kaarten. 1e-19e afl. Elke afl. f 0,50.
```

(b) OCR output of the catalogue snippet in (a)

Figure 3: A snippet from the 1833-1849 catalogue, displaying horizontal lines in the scan and mistakes in the OCR output

Henceforth, these volumes will be referred to by their starting year only, thus 1936-1940 becomes 1936, 1850-1882 will become 1850 etc.

4.2 Generate letter sections

The original printed books and PDF scans have the bibliographical entries separated by having the first word of each bibliographical entry in bold. However, this layout is not transferred to the OCR output, which poses a challenge when trying to find the beginning of an entry. The OCR output is a long file of lines without any indication of which lines should be grouped together as a single bibliographical entry. Therefore the lines the lines in the OCR output should be concatenated together to from the original bibliographical entries.

The bibliographical entries are listed alphabetically in the volumes, with headers, varying per volume, that indicate the scope of the bibliographical entries on a page. As a starting point to correctly group the lines together, the headers of a page were used while working on the 1980 volume. In 1980 the even-numbered pages have the first word of the page in the header, while the headers of the odd-numbered pages contain the final word of the page. The words in the headers of a page were used to create a range within which the first words of the bibliographical entries on certain pages should be. With some manual adaptations to the headings, this method was useful for the 1980 volume. Nonetheless, this method was not applicable to other volumes due to differences in the headers as well as the OCR quality of the headers for most volumes. This led to the realisation that a fully automatic system might not be reachable. Therefore, a manual pre-processing step is required. Even though this is manual labour and automatic processing is preferred, this method requires less manual annotation than adapting all the headers to fit the previous method.

The step consists of manually annotating the beginning of each letter in the OCR output of each volume. To prepare the data for this process, the lines containing headers are automatically removed by looking for lines containing only capital letters and non-alphabetic characters. The reason behind this is that the headers are not used for the creation of the bibliographical entries and would otherwise end up in an entry, which would lead to incorrect entries.

Using the PDF scans of the original printed books, the beginning of each letter was sought out and ‘START *’ was added at the corresponding place

in the OCR output for all volumes, where '*' is the capital version of the letter, see Figure 5b.

Unfortunately, transitions between letters within a page, rather than starting at the top of a new page, can cause complications. An example of such a transition between letters is shown in Figure 4. There are three possible main scenarios when this occurs which can be understood by looking at the figure.

1. The OCR process has correctly recognised the letter transition and groups the columns in the correct manner. This means that the order of the columns in the OCR output is 1-2-3-4.

2. The transition between letters is missed and the left column is placed before the right column in its entirety. This scenario corresponds to a 1-3-2-4 order of the columns.

3. The letter transition causes confusion and for one or both letters the two columns are concatenated into one. Putting the text from the right column directly after the left one's for each line. This implies that the lines in 1 and 2 are joined across the columns to form one line containing information from both columns. The same may occur for 3 and 4, or all four columns.

Of these three main scenarios the first one does not cause any problems and no action needs to be taken, since the OCR process recognises the columns correctly. For the other two scenarios problems arise because the lines, and therefore the bibliographical entries, are no longer in the correct order. The issue created by the second scenario is fixed by manually moving the incorrect column part to its right location in the file. The final scenario creates more difficulties as the lines of the two columns are concatenated. In cases where it concerned less than 20 lines this was manually corrected, for more lines it was left untouched. As a consequence items surrounding the letter transition will form incorrect bibliographical entries, though this hardly ever occurs on such a scale that it is not manually corrected. An example of this last scenario can be seen in Figure 5.

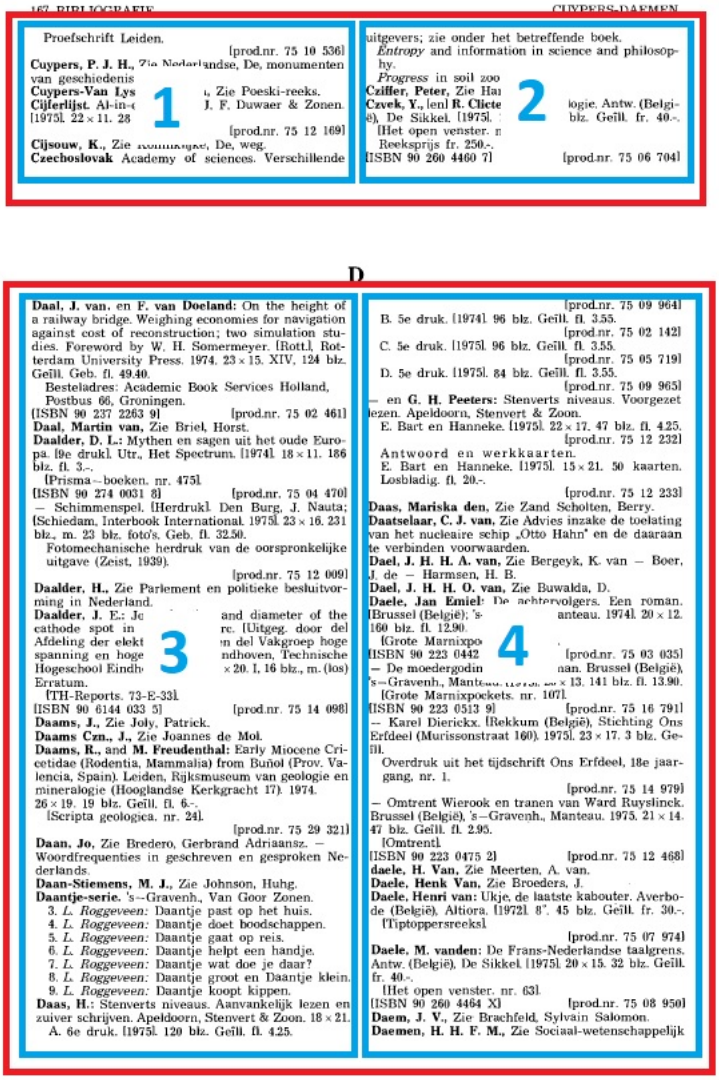

Figure 4: A letter transition within a page, where the columns that belong together are within a red rectangle and the four numbered columns parts in blue rectangles

```
Azamat Batuk [N. G. Thieblin] Spanje en Azeglio Roberto d',. Hot hof van Rome en
    hot Evangelie. eene stem nit Italic Gedr. voor
    rekening van den vertaler. Amst.. Gebr. Binger.
de  Spanjaarden.    Naar hot Eng. bewwerkt door
lo. de Vries.   Haarl., Kruseman & Tjeenk
Willink. (H. D. Tjeenk Willink. gr. 8o. in. (Gebr. Koster,. 1860. post So. f 0,35.
        Az-Zamaksarii Lexicon geographicum (Ara-
        bice). E cod. Leyd. nuns primum edidit M.
        Salverda de Grave. Lugd. Bat., E. J. Brill
        1856. 8o. maj f225.
portr. f3.25.
Azeglio (Masstmo d'). Mijne herinneringen.
Naar    de 6e uitg., uit het Italiaansch vert. door
H.
    J. Wansink. Kamp..  Laurens van Hulst.
1876. 3 dln. post 8o. f5.75.

B.
Baak ,E;, z;e Vennootschappen.
```

(a) Transition between letters where the columns are not recognised correctly.

```
Azamat Batuk [N. G. Thieblin] Spanje en
de  Spanjaarden.    Naar hot Eng. bewwerkt door
lo. de Vries.   Haarl., Kruseman & Tjeenk
Willink. (H. D. Tjeenk Willink. gr. 8o. in.
portr. f3.25.

Azeglio (Masstmo d'). Mijne herinneringen.
Naar de 6e uitg., uit het Italiaansch vert. door
J. Wansink. Kamp..  Laurens van Hulst.
1876. 3 dln. post 8o. f5.75.

Azeglio Roberto d',. Hot hof van Rome en
hot Evangelie. eene stem nit Italic Gedr. voor
rekening van den vertaler. Amst.. Gebr. Binger.
(Gebr. Koster,. 1860. post So. f 0,35.

Az-Zamaksarii Lexicon geographicum (Ara-
bice). E cod. Leyd. nuns primum edidit M.
Salverda de Grave. Lugd. Bat., E. J. Brill
1856. 8o. maj f225.

START B
Baak ,E;, z;e Vennootschappen.
```

(b) The corrected transition between letters with the columns in the correct manner.

Figure 5: The transition between letters corresponding to scenario 3.

4.3 Forming bibliographical entries by identifying the start position

In this section it is described how the bibliographical entries are be formed. To accomplish this certain lines from the OCR output are concatenated to form the entries after running multiple loops over the data.

At this point, some additional notes need to be made about the OCR output. They concern both minor manual adaptations to the OCR output and notes on the alphabetical ordering in the volumes.

- Some of the entries that should start with the letter 'X', mainly the entries that are written in Greek capital letters, do not start with an 'X' due to OCR errors. These letters have been manually changed to an 'X' to reduce the loss of bibliographical entries.
- The alphabetical order with respect to the Dutch vowel 'ij', in some volumes it is placed with the 'i' and in others it is treated as a 'y'.
  In case that it is treated as a 'y', the 'ij' is temporarily changed to a 'y'. If this were not done all the bibliographical items that start with a 'ij' are concatenated since they do not start with a 'y' and therefore are considered as a continuation of the previous entry rather than the start of a new one.
- *Titles* that start with a number such as '8 speciale onderwerpen voor het eindexamen' are listed alphabetically as if they were written in words, therefore this *title* is listed at 'acht'.
  To solve this issue a Python library called 'Telwoord' was used to replace the form written in numbers by the from written in words.. This needs to be done to avoid that the grouping algorithm misses all bibliographical entries that start with a number, because they do not start with a letter. By temporarily changing the number into its written form, we can get around this issue. This is only done temporarily because otherwise the *title* of the bibliographical entry would be changed and it would no longer correspond with the *title* in the volume.

4.3.1 Forming bibliographical entries based on first letter

To correctly group the bibliographical entries using the information in a line a few rules were created. Not all rules apply to each of the three groups of

volumes, following each rule is a list of the groups it applies to. By applying the rules a file with bibliographical entries is created for each volume. The first word is taken by splitting the line on spaces and taking the first item. This means that punctuation 'attached' to the first word is not removed immediately.

1. The indication of the start of each letter section is used to track the current letter in the alphabetical order. All bibliographical entries that are created should start with the capital version of the current letter, and cannot end with a full stop.
   All other lines are concatenated with the bibliographical entry until a new one is created. (All groups)

2. An apostrophe on the left side of the first word should be removed temporarily, unless the 'word' is the character is by itself. This rule is applied for *titles* that, for example, start with 'K. (All groups)

3. If a comma is present 'in' the first word, remove it from the word, unless the 'word' itself is a comma. (All groups)

4. Fix OCR mistakes by replacing a 0 at the beginning of the first word with an 'O', if this is the current letter. Do the same with replacing a 1 as first letter with an 'L', if this is the current letter. (All groups)

5. In the rare case that an entry starts with an initial that is corresponding with the current letter, remove the full stop to treat it as the start of a bibliographical entry. (All groups)

6. If the first word contains a dash, and it is not the first character of the word, temporarily remove it. (All groups)
   If it is the first character and there are indicators that the current bibliographical entry is complete, start a new entry. (Group 2 & 3)

7. If the first word contains a 'ij', replace it with a 'y'. (Group 1 & 2)
   Only replace it if the volume is not from the 19th century, 1850 and 1882. (Group 3)

By merging multiple lines of one bibliographical entry, a single line is formed per bibliographical entry. For the majority of the bibliographical

entries this needs to be done since they run over multiple lines due to space limitations.

However, for bibliographical entries that are describing series this is undesirable, because for those bibliographical entries the different parts of the series are described on different lines. An example of this can be seen in the penultimate entry of the third column in Figure 4 (Daantje-serie).

4.3.2 Dealing with discontinuous alphabetical order

The list of bibliographical entries is not alphabetical yet, because only the first letter of the word is used as a criteria. This results in errors were the alphabetical order of the bibliographical entries is discontinued. An example of this is shown in Figure 6, where the items that start with 'Alkmaar' considered bibliographical entries, however they belong to the bibliographical entries above.

```
Abkoude, Chr. van - Pietje Bell in Amerika / door Chr. van Abkoude ;
geheel opnieuw bew. door W. N. van der Sluys ; geill. door G. van
Straaten. - 22e dr. -

Alkmaar Kluitman, 1979. - 158 p : ill. ; 18 cm. - (Kluitman jeugdserie
; J 1081) (Pietje Bell serie) ISBN 90-206-1081-3 : f. 3.25 8030302

Abkoude, Chr. van - Pietje Bell in Amerika / door Chr. van Abkoude ;
geheel opnieuw bew. door W. N. van der Sluys ; geill. door G. van
Straaten. - 22e dr. -

Alkmaar : Kluitman, 1979. - 158 p : ill. ; 18 cm. - (Kluitman
jeugdserie ; J 1081) (Pietje Bell serie) ISBN 90-206-1081-3 : f. 3.25
8030302

Abkoude, Chr. van - Pietje Bell is weer aan de gang / door Chr. van
Abkoude. - 24e dr. - Alkmaar Kluitman, 1979. - 160 p. : ill. ; 18 cm.
- (Kluitman jeugdserie ; J 1042) (Pietje Bell serie) ISBN
90-206-1042-2 : f. 3.25 8030299
```

Figure 6: Bibliographical entries with a discontinued alphabetical order.

To improve the current list of bibliographical entries a second loop of grouping lines was used to pick up on these situations and to correct them. This loop was guided by the following algorithm:

Variables: current, prev, next, line, prev_line, next_line, new_line
Result: New combined lines where needed

new_line = line
if *next_line starts with '-'* **then**
  continue
else if *(current < prev* ***and*** *current < next)* ***or*** *(current > prev* ***and*** *current > next)* **then**
  new_line = prev_line + line
  delete prev_line
  delete line
else
  continue
end

Algorithm 1: Combine lines where the alphabetic order is broken

The algorithm has multiple conditions that determine whether two lines should be concatenated or not. These conditions are for the most part based on the first word of the current line, the previous line and the next line. The alphabetical order of two of these words is checked at a time. An exception is found if the next line starts with a dash for repeated information, in that case lines are never concatenated.

If the current first word is earlier in the alphabet than both the previous first word and the next first word, the previous line and the current line are concatenated to form one new line. The original previous and current lines are then deleted. The same process holds when the current first word is later in the alphabet than both the previous first word and the next first word.

The result of running the data from Figure 6 through the algorithm can be seen in Figure 7. The incorrect bibliographical entries that started with 'Alkmaar' have correctly been combined with the previous bibliographical entry to restore the alphabetic order of the items.

4.3.3 Merging remaining incorrect entries

As mentioned in Section 4.3.2 some lines are not properly combined due to the start letter of the line. Most cases are taken care of by the method described in the section mentioned above. Nonetheless, some lines are still concatenated incorrectly because they happen to start with the same word

```
Abkoude, Chr. van - Pietje Bell in Amerika / door Chr. van Abkoude ;
geheel opnieuw bew. door W. N. van der Sluys ; geill. door G. van
Straaten. - 22e dr. - Alkmaar Kluitman, 1979. - 158 p : ill. ; 18 cm.
- (Kluitman jeugdserie ; J 1081) (Pietje Bell serie) ISBN
90-206-1081-3 : f. 3.25 8030302

Abkoude, Chr. van - Pietje Bell in Amerika / door Chr. van Abkoude ;
geheel opnieuw bew. door W. N. van der Sluys ; geill. door G. van
Straaten. - 22e dr. - Alkmaar : Kluitman, 1979. - 158 p : ill. ; 18
cm. - (Kluitman jeugdserie ; J 1081) (Pietje Bell serie) ISBN
90-206-1081-3 : f. 3.25 8030302

Abkoude, Chr. van - Pietje Bell is weer aan de gang / door Chr. van
Abkoude. - 24e dr. - Alkmaar Kluitman, 1979. - 160 p. : ill. ; 18 cm.
- (Kluitman jeugdserie ; J 1042) (Pietje Bell serie) ISBN
90-206-1042-2 : f. 3.25 8030299
```

Figure 7: Bibliographical entries where the alphabetical order has been restored.

or name as the start of the bibliographical entry. The example in (2) shows two bibliographical entries that should be one entry, but they are ordered correctly alphabetically and hence are not concatenated by the method in Section 4.3.2.

To identify the cases where this issue occurs a loop is created that checks both the first word of the current bibliographical entry as well as the first word of the previous one. If they are identical and the second 'word' in the current bibliographical entry is a semicolon, the two entries are concatenated to create one complete entry.

(2) Abkoude, Chr. van - Kruimeltje / door Chr. van

Abkoude ; [ill. Pol Dom]. - 46e dr. - Alkmaar : Kluitman, [1980]. - 168 p. : ill. ; 23 cm. - (Populaire jeugdboeken) ISBN 90-206-2004-5 geb. : f. 9.95 8030341

4.3.4 Replacing the dashes

The final step of this phase is replacing the dashes, due to repeated information as mentioned in Section 4.1, with the correct content. To replace the dashes with the correct information, either an *author* or the first word of

a *title*, it has to be determined whether the previous bibliographical entries starts with an *author* or a *title*.

To accomplish this the bibliographical entries are split, following the method described in Section 4.4.1. The current *author* and *title* are saved before continuing with the next bibliographical entry. If a bibliographical entry starts with a dash, the *author* and *title* of the previous bibliographical entry are checked. In case the previous bibliographical entry has an *author*, the *author* is inserted replacing the dash, otherwise the first word of the *title* of the previous bibliographical entry replaces the dash. Examples where the *author* is the replacement and the first word of the *title* can be seen in Figure 8 and 9 respectively. Replacing the dashes reduces the difficulty to extract the *author* and *title* in the next stage. A downside is that the extraction technique does not always recognise an author after the dash, if a bibliographical entry has multiple *authors*, which can cause an incorrect representation of the *authors* of the bibliographical entry.

```
Aalders, W. J.: De analogia entis in het geding. Zie Mededeel

-- De apostolische geloofsbelijdenis. Zeist, J. Ploegsma. 193

  - De grond der zedelijkheid. (Wolters). Panic. prijs opgehe
```

(a) Dashes in the bibliographical entries instead of the *author*.

```
Aalders, W. J.: De analogia entis in het geding. Zie Mededeel

Aalders, W. J.:  De apostolische geloofsbelijdenis. Zeist, J.

Aalders, W. J.:  De grond der zedelijkheid. (Wolters). Panic.
```

(b) Dashes are replaced by the correct *author*.

Figure 8: The dashes in the bibliographical entries are replaced with the corresponding *author*.

4.4 Extracting specific metadata from a book entry

To extract the bibliographical metadata from the entries, the use of different text mining techniques is explored. This includes a rule-based approach (4.4.1) as well as Named Entity Recognition (4.4.3) and a Probabilistic

```
Administratle, De, van bet b a k k e r s-b e d r ij f. Leiden
- De,voor bouwondernemers. Leiden, Ned. uitg.bedr. van wetens
- De, van een c a f e b e d r ij f. Lei-den, Ned. uitg.bedr.
- De, van manufacturenwin-k e 1 s. Leiden, Handelswetenschapp
```

(a) Dashes in the bibliographical entries instead of the first word of the *title*.

```
Administratle, De, van bet b a k k e r s-b e d r ij f. Leiden
Administratle,  De,voor bouwondernemers. Leiden, Ned. uitg.be
Administratle,  De, van een c a f e b e d r ij f. Lei-den, Ne
Administratle,  De, van manufacturenwin-k e 1 s. Leiden, Hand
```

(b) Dashes are replaced by the first word of the *title*.

Figure 9: The dashes in the bibliographical entries are replaced with the correct word from the *title*.

Context-Free Grammar (4.4.2). The metadata that should be extracted from the bibliographical entries include the *author*, *title*, *city_of_publication*, *publisher*, *year_of_publication*, *size_of_book*, *price_of_book*, *ISBN_number*, *number_of_pages* and a (Brinkman) *production_ number*. If a technique is unable to extract a piece of metadata, whether it is not present in the entry or missed by the extraction technique, it is set to 'Unknown'.

4.4.1 Rule-based with regular expressions (RB-REGX)

In the RB-REGX technique the metadata from the bibliographical entries is extracted using rules and regular expressions for all nine volumes of the catalogue. Since the volumes are grouped together based on their layout and structure, the three groups require a different set of rules. For each group the extraction is based on the structure of the bibliographical entry, more specifically whether there is a marker, such as punctuation, that denotes the separation between certain pieces of metadata.

Elements such as the *ISBN_number*, *price_of_book* and Brinkman *production_number* can be extracted by regular expressions. Due to their consistent format they rely less on the order of other metadata within the bibliograph-

ical entry and can be extracted without the use of other information.

Some metadata is only sporadically given in the bibliographical entry and inconsistently when it comes to notation, such as which print of the book it concerns. These pieces of information are currently not extracted.

4.4.1.1 Extracting information: Group 1

The group containing the most recent volumes included in my data set, 1979 and 1980, has clear separations between the different pieces of metadata. An example of a bibliographical entry from 1980 is shown in Figure 10. The punctuation displayed in red is the punctuation that is used to extract the metadata using sentence splitting and determining what kind of metadata it is.

Bibliographical entries start with either a *title* or an *author* and a *title*, using this information the bibliographical entry is initially split on the first '-'-character. It is then determined whether this part of the bibliographical entry concerns an *author* or whether it includes the *title* up to the publishing information. For this part to be recognised as an *author* it has to have a length of less than 75 characters, contain either exactly one comma or indications of a pseudonym, and it cannot contain a forward slash. In case an *author* has been identified, the part of the bibliographical entry after the first '-' is considered next. If not, a *title* is extracted from the part up to the first '-' by splitting the bibliographical entry on the forward slash, if present.

After identifying either an *author*, *title* or both, the section of the bibliographical entry after the first dash is processed. This part is split on all the dashes, resulting in multiple sections that are divided by a dash in the full bibliographical entry. The program loops through these sections looking for one containing a colon, which indicates the presence of the publishing information that should be extracted.

The publishing information that is part of the bibliographical entry contains three different pieces of information: the *city_of_publication*, the *publisher* and the *year_of_publication* of the bibliographical entry. The *city_of_publication* is extracted by taking the first part after splitting the section on the colon. The remaining part holds the *publisher* and *year_of_publication*, they are extracted by splitting on a comma. On the left of the comma the *publisher* can be found, while for the right-hand side a regular expression is used to extract the *year_of_publication*.

Next the part of the bibliographical entry that is left over is split on a

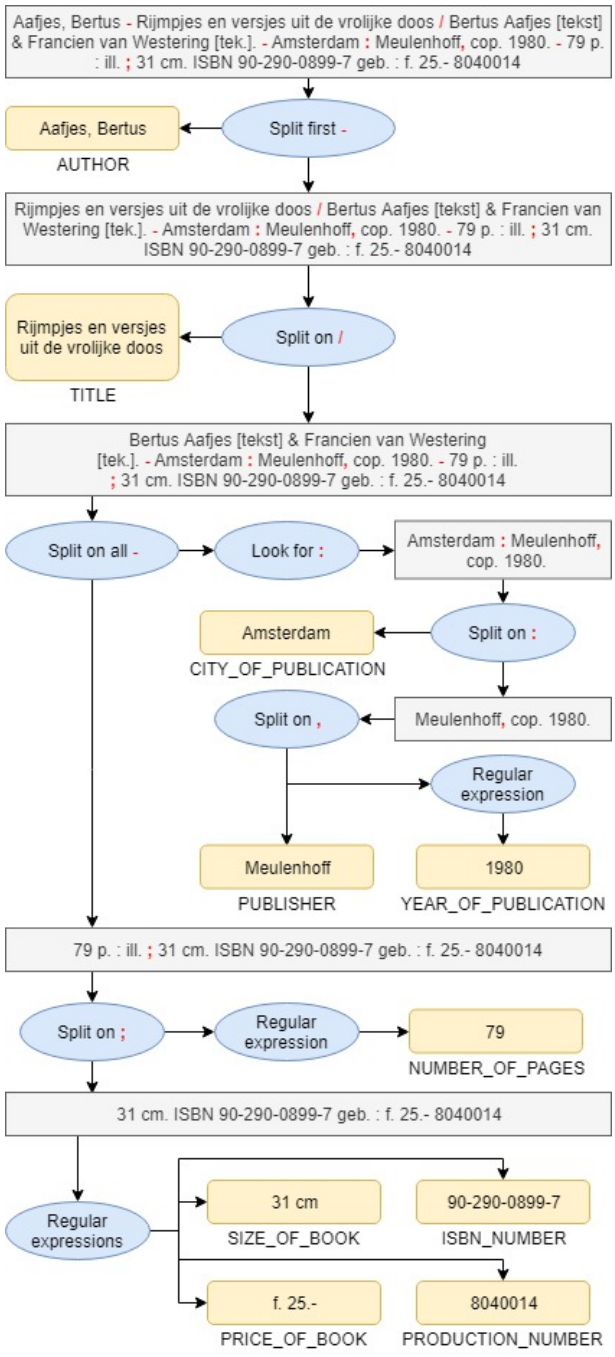

Figure 10: Process of extracting bibliographical metadata form a bibliographical entry from a volume in group 1.

semicolon, to easily extract the *number_of_pages* using a regular expression. The remainder of the bibliographical entry is questioned by multiple regular expressions in order to extract the *size_of_book*, *ISBN_number*, *price_of_book* and *product_number* of the bibliographical entry. If multiple prices are found in the bibliographical entry only the last one is extracted, this is the newest price.

4.4.1.2 Extracting information: Group 2

Extracting information from the second group of volumes is harder than the first. The reason for which is that unlike in the volumes of the first group, the volumes in the second group do not use many different punctuation markers to indicate the boundaries of information. With the exception of the *author*, and a comma between the *city_of_publication* and *publisher*, all pieces of information end with a full stop. A flowchart of the extraction process for a bibliographical entry from the 1961 volume can be seen in Figure 11.

Because nearly all the information is separated by a full stop it is hard to correctly extract the information. For example, a *city_of_publication* can be written with an abbreviation such as 'Amst.' which brings a full stop into the bibliographical entry that is not the separation for this piece of information. The *city_of_publication* and *publisher* are separated by a comma, which may be missed due to the full stop from the abbreviation. Additionally, *titles* can include names with initials, and the *author* may also include initials followed by full stops. Due to this undesirable effect, using the full stops to find the separation between pieces of information is highly variable in number of full stops and unpredictable. A *title* that contains three full stops, for for example initials, will give the impression that we have three different pieces of metadata when extracting using the full stop. Hence, the extracted metadata, such as th *city_of_publication* and *publisher*, are likely to be incorrect for a large number of bibliographical entries.

The *author* is extracted by splitting the bibliographical entry on the first colon. The remaining section of the bibliographical entry is then split on the first full stop which is taken as the *title*, possibly incorrect if initials are included in the *title*.

If a comma is present in the remaining part of the entry, the entry is split on this comma. The *city_of_publication* is extracted by taking the last element before the comma, while the *publisher* is the first element after the comma. Elements are separated by full stops, this means that a *publisher*

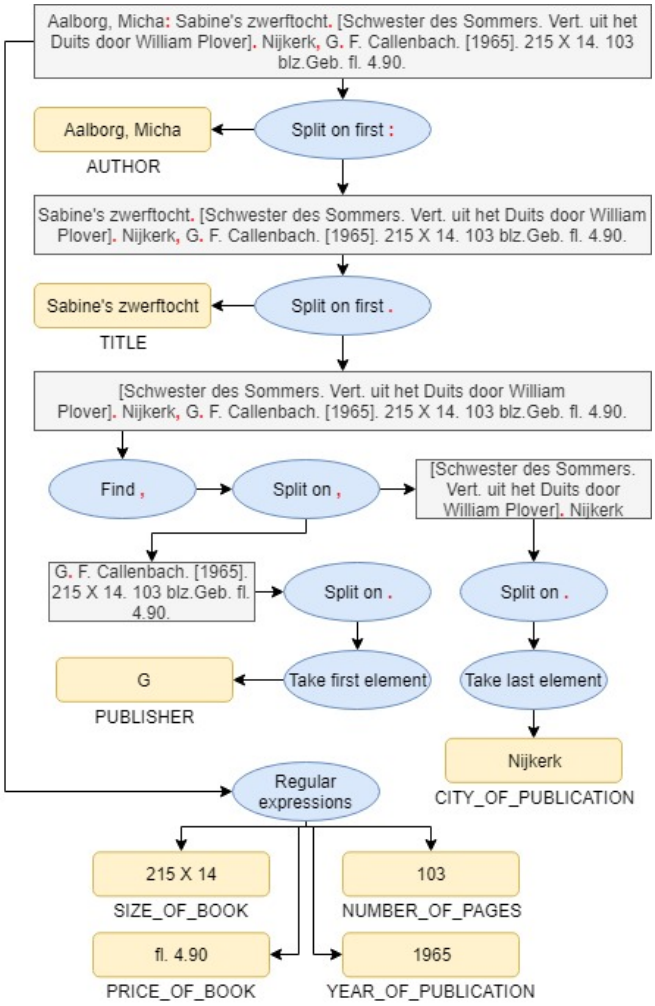

Figure 11: Process of extracting bibliographical metadata form a bibliographical entry from a volume in group 2.

that has initials in the name will be cut short.

Regular expressions are used to extract the *price_of_book*, *size_of_book*, *year_of_publication* and *number_ of_pages*.

For the 1971 and 1975 volumes of this group the *ISBN_number* and *production_number* are also extracted by the means of regular expressions. These pieces of bibliographical metadata are not present in the other two volumes of the second group.

4.4.1.3 Extracting information: Group 3

The oldest volumes of the catalogue contain the least information, as compared to the more recent volumes in the other groups. They also contain relatively many OCR mistakes which complicates the extraction process. The process of extracting information from a bibliographical entry from the 1882 volume is shown in Figure 12.

The volumes in the third group, like those in the second group, contain almost solely full stops as separators between pieces of information. This causes the same issues as mentioned before, where finding separations is sometimes impossible.

The *author* is extracted by splitting on the first comma of the bibliographical entry. For this part to qualify as an *author* it need to have rounded brackets, initials are between them, and a length of less than 75 characters. Unfortunately, some parts of *titles* qualify as well such as 'Almanak (Deventer Hoveniers-), 1851-1883.'. For this example 'Almanak (Deventer Hoveniers-)' is taken as the *author* of the bibliographical entry.

The *title* of the bibliographical entry is extracted by splitting the bibliographical entry on either the first full stop after the *author*, if present, or the first full stop of the bibliographical entry.

If a comma is present in the remaining part of the entry, the entry is split on this comma. The *city_of_publication* is extracted by taking the last element before the comma, while the *publisher* is the first element after the comma. Elements are separated by full stops, this means that a *publisher* that has initials in the name will be cut short.

Other pieces of bibliographical metadata including the *city_of_publication*, *publisher*, *year_of_publication*, *size_of_book* and *price_of_book* are extracted using regular expressions. For 1931 the *number_of_pages* are also extracted using regular expressions, the metadata concerning the *number_of_pages* is not present in both volumes from the 19th century.

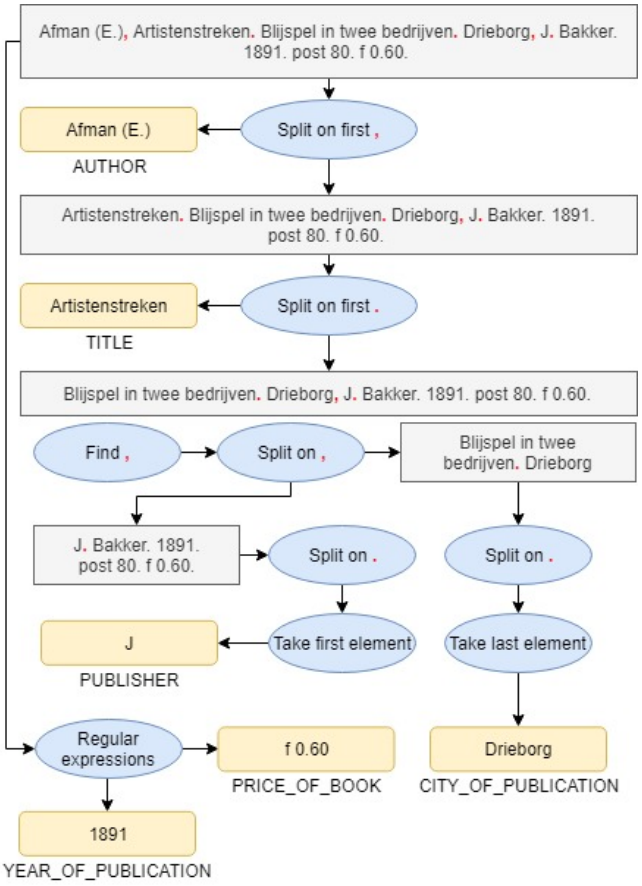

Figure 12: Process of extracting bibliographical metadata form a bibliographical entry from a volume in group 3.

4.4.2 Probabilistic Context-Free Grammar (PCFG)

Two Probabilistic Context Free Grammars are created to extract information from the bibliographical entries for the 1971 and 1980 volumes. Due to time constraints this technique is only applied to these two volumes.

The grammars consist of production rules and a lexicon, both have probabilities attached to them based on how often they occur. To get an approximate of the correct probabilities for the production rules a sample of 10 random bibliographical entries is taken and studied. Based on these 10 bibliographical entries the rules are created and the corresponding probabilities for the rules are calculated.

The lexicon is created by tokenising all the bibliographical entries in a volume and a set of unique tokens is made. The tokens are classified into four different groups: names, numbers, punctuation and other words. A token is classified as a name when it starts with a capital letter, a number is a token that contains at least one number and no letters. The punctuation is taken separately to exclude them from the automatic creation of the lexicon, they have to be manually defined to express boundaries between information in production rules. All other tokens are classified as words. For the names, numbers and words the probability of each token is then calculated by taking $\frac{1}{no\_tokens\_in\_set}$, the probability is forced to a number with ten digits behind the decimal point, fully written out. This is done to avoid the scientific notation of these small numbers, because the scientific notation causes an error since it is not fully numeric. To parse the bibliographical entries the first 100 tokens of every bibliographical entry are obtained. The limit of 100 tokens is set to avoid dealing with extremely long bibliographical entries that the parser cannot parse, as well as increase the efficiency of the parser itself.

The main benefit of using the PCFG technique compared to the RB-REGX technique is that it can deal with minor OCR mistakes. Given that a *city_of_publication* and *publisher* are normally notated as: *city_of_publication*, *publisher*. It can be that the comma separating the two pieces of information is accidentally a full stop due to an OCR error. By using the probabilities it can still be recognised as a publishing sequence. For example, comma can get a 95% of being an actual comma and a 5% chance of being a full stop. A disadvantage, however, involves the effort needed to create the grammar itself. The grammar needs to cover all the possibilities, if the parser comes across an unknown sequence (perhaps due to an OCR mistake) it will not be able to parse the bibliographical entry. There is no option available to

```
bibliographical_entry -> title slash extra_info dash publish_info comma year dash pages semicolon size extra_info ISBN
extra_info price prod_nr [0.25]
bibliographical_entry -> title slash extra_info dash publish_info comma year dash pages semicolon size extra_info price
prod_nr [0.25]
bibliographical_entry -> title slash extra_info dash publish_info comma year dash pages semicolon size extra_info
prod_nr [0.2]
bibliographical_entry -> title slash extra_info [0.05]
bibliographical_entry -> author dash bibliographical_entry [0.25]
substring -> N [0.2] | Name [0.2] | comma [0.1] | dash [0.05] | period [0.05] | number [0.1] | semicolon [0.1] | colon
[0.1] | bracket [0.1]
author -> Name comma Name [0.6] | Name comma [0.2] | author substring [0.2]
title -> Name [0.1] | title substring [0.9]
extra_info -> substring [0.1] | extra_info substring [0.9]
print -> N N period [0.6] | N N [0.2] | bracket print bracket [0.1] | print period [0.1]
publish_info -> city colon publisher [1.0]
city -> Name [0.5] | substring city [0.2] | city substring [0.2] | city semicolon city [0.1]
publisher -> Name [0.4] | substring publisher [0.2] | publisher substring [0.2] | Name dash Name [0.1] | publisher Name
dash Name [0.1]
year -> number period [0.6] | substring year period [0.1] | substring period number period [0.1] | bracket number
bracket period [0.2]
pages -> number N [0.6] | bracket number bracket N [0.1] | pages colon extra_info [0.2] | pages comma extra_info [0.1]
size -> number N [0.2] | number N period [0.8]
ISBN -> "ISBN" isbn_substring [0.3] | ISBN isbn_substring [0.7]
isbn_substring -> number [0.4] | dash [0.4] | N [0.2]
price -> N period number [0.7] | "f." number [0.2] | N number [0.05] | N N [0.05]
prod_nr -> number [1.0]
bracket -> "[" [0.25] | "]" [0.25] | "(" [0.25] | ")" [0.25]
comma -> "," [1.0]
dash -> "-" [1.0]
slash -> "/" [1.0]
```

Figure 13: Part of the grammar for the 1980 volume, each production rule describes a possible structure of an element with probabilities attached to it.

```
(bibliographical_entry
  (author (Name Abma) (comma ,) (Name Willem))
  (dash -)
  (bibliographical_entry
    (title
      (title (title (Name En)) (substring (N it)))
      (substring (N barde)))
    (slash /)
    (extra_info
      (extra_info
        (extra_info (substring (Name Willem)))
        (substring (Name Abma)))
      (substring (period .)))
    (dash -)
    (publish_info
      (city_of_publication
        (substring (Name De))
        (city_of_publication
          (substring (Name Jouwer))
          (city_of_publication
            (city_of_publication (substring (N [)) (city_of_publication (Name Joure)))
            (substring (N ])))))
      (colon :)
      (publisher (Name Hynsteblom)))
    (comma ,)
    (year_of_publication (number 1980) (period .))
    (dash -)
    (number_of_pages (bracket [) (number 28) (bracket ]) (N p.))
    (semicolon ;)
    (size_of_book (number 18) (N cm) (period .))
    (extra_info
      (extra_info (substring (N f.)))
      (substring (N 15.-)))
    (production_number (number 8070804))))
```

Figure 14: The parse tree created by the PCFG for a bibliographical entry

partially parse a bibliographical entry, it is either parsed completely or not at all.

An example of a parse tree created by the PCFG is shown in Figure 14.

To be able utilize the data extracted from the bibliographical entry by the grammar in the following step, the information from the parse tree needs to be saved in an accessible format. To extract the information from the parse tree the number of opening and closing brackets is counted.

The counter for the opening brackets increases by one every time an opening bracket is encountered. When a closing bracket is found the number of opening brackets is decreased by one. Because the main 'bibliographical entry' is not closed until the very end of the tree, the current piece of

information is finished when the opening brackets counter equals one.

Because recursion is used to simplify the production rules, all the rules starting with a *title* can come after an *author*, the 'bibliographical entry' branch is opened twice. To ensure that second 'bibliographical entry' is not taken as an item name, a counter is implemented to keep track of the number of 'bibliographical entry' openings it comes across. When there is more than one, the second one is ignored.

Apart from the first line, all the information is concatenated as long as the opening branch counter is higher than one. As soon as it reaches one the item is saved and a new item is started. This way all the pieces of information are saved as individual items. However, they still include all the branch names and not the content itself. To solve this, the items are split on the space and only the parts containing a closing bracket, the end of an item or sub item and thus content, are saved with their type after being stripped of the closing bracket.

The information is then transferred to a file in the same format as the output of the RB-REGX technique. This makes it easier to evaluate the information using the same methods as for the output of the RB-REGX technique.

4.4.3 Named Entity Recognition

The Named Entity Recognition technique is used to extract the some of the named entities from the bibliographical entries. The main objective for applying this technique is to obtain the *publisher* of a book, since the position of this information in an item can vary and the it often consists of multiple words.

To extract the named entities from the bibliographical entries, the Dutch language model of Spacy [6] is used. Spacy classifies the named entities into one of the following four categories: PER (person), MISC (miscellaneous), LOC (location) or ORG (organisation). The examples below (3)-(6), show that recognition of named entities is not always correct, both with extracting the entities as well as with classifying them.

(3) Hasman, Arie: Neutron quasi-elastic scattering studies on fluid argon. [Met een samenvatting in het Nederlands]. Rott., Bronder-offset (Goudsesingel 260). 1971. 23 x 16. 123 blz., m. (los) errata. Geill.

[6] https://spacy.io/models/nl

Proefschrift Delft. [prod.nr. 7144055]

Recognised entities:
Hasman (PER), Arie (PER), Neutron quasi (PER), Rott (PER), Bronder (LOC), Goudsesingel (LOC), Geill (PER) and Proefschrift Delft (PER)

Example (3) demonstrates that a *publisher* that contains a hyphen is not recognised as a single entity. From the *publisher* in the example 'Bronder-offset', only 'Bronder' is recognised as an entity, while 'offset' is not recognised as a separate entity either.

(4) *Examples of recognised publisher entities:*
Nederlandse Boekenclub (ORG), Rijksuniversiteit (LOC), J. F. Duwaer & Zonen (PER) and A. Asher & Co (MISC)

(5) Meulenberg, M. T. G.: Inleiding tot de marktkunde. Utr.; Antw. (België), Het Spectrum. [19711. 18 x 11. 213 blz. fl. 6.-. [Marka-boeken. nr. 1121. [ISBN 90 274 6093 OJ [prod.nr. 71220561

Entity 'Het Spectrum':
Het Spectrum (LOC)

(6) Piepenstock, Marianne: De franse keuken. [Französische Küche. Vert. uit het Duits door A. Hol druk. Utr., Het Spectrum. [1971].-laar-Pitstra]. 4e 18 x 11. 143 blz. fl. 2.25. [Prisma-boeken. nr. 13781 [ISBN 90 274 0388 0] [prod.nr. 7147156]

Entity 'Het Spectrum':
Het Spectrum (MISC)

The examples (4)-(6) reveal that recognised *publisher* entities are classified into different categories which makes the process of recognising it as a *publisher* hard. It can particularly be seen in (5) and (6), in which the same *publisher* 'Het Spectrum' is classified in different categories.

Probably, these inconsistencies can be explained by the nature of the current data. The bibliographical entries are not running text, which is what the Dutch Spacy language model is trained on.

As a result, this technique cannot be used as a method to extract the *publisher*, or any other piece of metadata. In Section 4.5.2, however, I will explain how Named Entity Recognition is used to improve the extracted metadata.

4.5 Using external knowledge

External knowledge sources are used in the extracted bibliographical metadata in two different ways. Gazetteers are used to give the *author* and *city_of_publication* a confidence score and NER is used to improve the extracted metadata by replacing an 'Unknown' *city_of_publication* and/or *publisher* with an appropriate named entity. In the evaluation phase the confidence scores will be used to see whether only including metadata with high confidence scores improve the results.

4.5.1 Gazetteers

Three different gazetteers are used:

- *Dutch cities*: This list contains the names of Dutch cities and is extracted from the Metapopos website [7]. It consists of 2423 cities located in The Netherlands. The list is supplemented with 'Den Haag' and 'Den Bosch', since only ''s-Gravenhage' and ''s-Hertogenbosch' are included in the list while both are found in the catalogue. Also, 'Amst', 'Rott' and ''s-Gravenh' are added, since they are common abbreviations in the catalogue.
- *Belgian cities*: The list of Belgian city names is extracted from a GitHub project [8]. It is based on the list of postal codes used by the Belgian Postal Services and was retrieved on September 11 2015. It contains a total of 2712 cities located in Belgium.
- *Dutch surnames*: This is a list of Dutch surnames [9], scraped from the Meertens Institute website by an employee of the National Library of the Netherlands. It consists of nearly 300,000 names, including 'variantions' of the same surname differentiated by affixes such as 'der' and

[7] http://www.metatopos.eu/almanak.html

[8] https://github.com/spatie/belgian-cities-geocoded

[9] https://github.com/WillemJan/Narralyzer_languagemodel/tree/master/pp/lang/nl

'den'. Capitalisation is also present in the list, both 'Berg, Van der' and 'Berg, van der' are included. Which means that when capital letters are lowered, the list does not contain 300,000 unique surnames.

These gazetteers are used to give the *author* and *city_of_publication* a confidence score between 0 and 2 or 'Unknown' in case the metadata is 'Unknown'.

4.5.1.1 Authors

To give the extracted *author* a confidence score the *author* is first tokenised. This is done to make it possible to match on the surname without affix, first name or initials. If this were not done an *author* like 'Lievense-Pelser, E.' would not be recognised, and even 'Lievense-Pelser' might not be present in the list. By tokenising the extracted *author* it checks both 'Lievense' and 'Pelser' against the list, thus having a higher chance of finding a match with a listed surname.

If the extracted *author* is found in the gazetteer is gets a confidence score of 2. In the case that the extracted *author* has the correct structure, for example a word starting with a capital letter followed by a comma and another word with a capital letter, it is given a confidence score of 1. In all other cases it is unlikely that the extracted *author* is indeed an *author* and the confidence score is set to 0. When the *author* is 'Unknown' the confidence score is also set to 'Unknown'.

4.5.1.2 Cities

The *cities_of_publication* are checked both tokenised as well as untouched, because some *cities_of_publication* are written between brackets or have additional information behind them, such as 'Amsterdam [etc.]', by tokenising it 'Amsterdam' can be recognised as a *city_of_publication*. However, due to the tokenisation process cities that consist of multiple tokens like 'Den Haag' are not recognised anymore. For this reason *cities_of_publication* are also checked as a whole to not miss any cities that would otherwise not be identified. When the *city_of_publication* is 'Unknown' the confidence score is also set to 'Unknown'.

If the *city_of_publication* is found in the gazetteer is gets a confidence score of 2. In the case that the *city_of_publication* has the correct structure, a word starting with a capital letter, it is given a confidence score of 1. In all other cases it is unlikely that the *city_of_publication* is indeed a *city_of_publication*

and the confidence score is set to 0.

4.5.2 Named Entity Recognition

The Named Entity Recognition appeared not to be an efficient and accurate technique to extract publishers details from the bibliographical entries, see Section 4.4.3. In this section, however, the extracted named entities are further processed to fill gaps in the extracted metadata, if the *city_of_publication* and/or the *publisher* is 'Unknown'. This means that this specific piece of metadata is not extracted from the bibliographical entry.

The metadata extracted using the RB-REGX and PCFG techniques are analysed to find the *city_of_publication* and/or the *publisher* where the value is 'Unknown'. For these bibliographical entries the Named Entity Recognition system is run to extract the named entities.

The unknown *city_of_publication* is replaced by the first Named Entity with the type 'LOC', if such an entity has been extracted. The type 'LOC' has been chosen because city names are locations and therefore the type 'LOC' (location) is most appropriate.

The first Named Entity with either 'ORG' or 'MISC' as type replaces the 'Unknown', to provide a name for a *publisher*. These two types, organisation and miscellaneous, are taken because they seem to match best with the *publisher*, as opposed to a person 'PER' or a location 'LOC'.

4.6 Conversion to PICA+

In order to easily link the extracted data with data already available in the library catalogue of the National Library and to make it available for further use by others, the extracted data is converted to the PICA+ format. This format consists of a list of fields for the metadata about the book. Separate fields are in place for the main author, second author, language of the book, size of the book, publication date, title, publisher etc. Not all the fields will be used for the output of the current research since not all the metadata is extracted and available.

004A $0*ISBN_number*$f*price_of_book*
006C $0*production_number*
011@ $a*year_of_publication*

021A $a@*title*
028A $a*author*
033A $p*city_of_publication*$n*publisher*
034D $a*number_of_pages*
034I $a*size_of_book*

Most of the metadata can be filled into this scheme the way they are extracted from the bibliographical entries. Exceptions to this are the *author* and *title*. For the *title* the first main word in the *title* should be preceded by a '@', this means that if the *title* starts with an article, the '@' has to be inserted after this word instead of at the beginning, see (7) and (8). The *author* has to be split into three parts, if all three are presents: the surname, the affixes such as 'van der' and the first name and initials. When all three elements are present the PICA+ format will be:

028A $d*firstname and/or initials*$c*affixes*$a*surname*

(7) *Title*: Koninklijke liefde
PICA+ notation: 21A $a@Koninklijke liefde

(8) *Title*: De leeuwentemmer
PICA+ notation: 021A $aDe @leeuwentemmer

In Appendix A a few complete bibliographical entries are given in the PICA+ format.

5 Evaluation

To be able to evaluate the extracted data a set of evaluation data needed to be created for the majority of the volumes. The evaluation data of the three most recent volumes, 1975, 1979 and 1980, is already digitally available. This data is extracted from the main catalogue of the National Library of The Netherlands and can be linked with the extracted data by means of a Brinkman production number.

A total of 100 bibliographical entries from each volume are taken to be evaluated and a comparison is made between the extracted data and the gold data in the evaluation sets. For the two volumes from which metadata is extracted using two different techniques, the same evaluation set is used to evaluate both techniques.

5.1 Creating manual evaluation data

For the six oldest volumes of the Brinkman catalogue used in this research there is no evaluation data digitally available. It cannot be guaranteed that the bibliographical entries entered in the library catalogue are identical to the ones described in the catalogue volumes due to the lack of a Brinkman production number. For this reason evaluation data is manually created for these volumes. Table 3 shows which pages were taken to create the evaluation data for each volume.

Volume	**Pages**
1850	68, 555, 598, 926 (partial)
1882	34, 85, 392, 565, 632 (partial)
1931	538, 900, 1074, 1086 (partial)
1936	*Part 1*: 119, 448, 595 *Part 2*: 374, 579, 619
1961	*Part 1*: 193, 1341 *Part 2*: 605, 1275
1971	17, 36, 165, 212, 425, 630 (partial), 701 (partial)

Table 3: Overview from which pages bibliographical entries are taken to create the evaluation data for each volume.

5.1.1 Guidelines for creating manual evaluation data

Some guidelines are enforced to ensure that the evaluation data is created in a consistent manner across the different volumes. Not all the information is copied, for example, the address of a *publisher* is skipped, only the relevant information is copied which is the main *publisher*. To get 100 bibliographical entries per volume, random pages from the volume are selected. The metadata from the book items on these pages are are copied while applying the following rules:

1. Bibliographical entries that run off the page or are not complete at the beginning are disregarded. Only entries that are fully on the particular page are copied.

2. Like in the extraction process all bibliographical entries containing a reference to another bibliographical entry are disregarded. Because no metadata is extracted from these bibliographical entries they do not need to be evaluated and therefore would only contaminate the evaluation data.

3. If the bibliographical entry is the main entry of a book series, only the main *title* of the book series is taken. The separate book titles of books in the series are nestled in such a way that it is extremely hard to extract those from the bibliographical entry. For this reason only the main *title* is extracted by the extraction techniques and should be evaluated. The *price* is in this case 'Unknown' since each book has its own price, this also holds for the *year_of_publication*, *ISBN_number* and *production_number*.

4. For bibliographical entries that start with a dash due to duplicated information, the correct metadata is copied by checking the previous bibliographical entry. This process of replacing the dashes is also done before extracting the metadata and avoids issues with missing information.

5. Sometimes words contain spaces in between the letters to emphasise them. For example, Almanak (Provinciale) voor Z u i d-H o l l a n d. For the evaluation data these spaces are removed, because the real title of the bibliographical entry does not contain these spaces and therefore it would not show the title of the entry but rather how it is displayed

in the catalogue volume. The OCR process also removes most of the spaces between the letters thus it causes no problem for the evaluation of the extracted data.

6. Metadata that have diacritics in them are taken as they are, and the diacritics are copied.

7. When multiple publishers are mentioned in a bibliographical entry, the first one is taken.

8. If a *publisher*'s address is mentioned, this information is skipped.

9. Bibliographical entries do not have a *price_of_book* but instead: 'Part. prijs opgeheven' (private price canceled), the *price_of_book* is copied as 'Unknown'.

5.2 Digitally available evaluation data

The evaluation data for the volumes 1975, 1979 and 1980 are already available in a digital form and can be extracted from the National Library catalogue. Using a specific number (Brinkman production number) these catalogue entries can be linked to the extracted entries. This number is an identification number within the Brinkman catalogue volumes and the National Library catalogue.

Since the National Library has only been actively acquiring all published books since 1974, very few links can be made for bibliographical entries prior to this date. Therefore, only the three most recent volumes qualify for this method of acquiring evaluation data.

The bibliographical entries that are selected for the evaluation data from the 1975, 1979 and 1980 volumes, need to have a Brinkman production number that can be linked with the digitally available evaluation data. Therefore the number should be extracted correctly and be present in the National Library catalogue. Because of this requirement the entries are not selected from random pages, rather a selection of 100 random entries are taken from the intersection between the Brinkman production numbers that are correctly extracted from the entries and the numbers that are present in the digital evaluation data. A downside to this evaluation method is that the aspect of 'completeness' (see below) cannot be evaluated.

5.3 Evaluation process

The extracted bibliographical metadata from the entries and the created entries themselves are evaluated on multiple aspects:

- *Completeness*: Not all entries are correctly found as some of them are (partly) concatenated with previous or following ones. Bibliographical entries might be missing from the formed entries, by means of being merged with another entry, and entries might be formed that are not originally an entry in the catalogue volume.
- *Exact versus fuzzy matches*: The data used in this research contains OCR errors. These errors are therefore also present in the extracted metadata. To account for these errors, the extracted metadata is evaluated using both exact matching and fuzzy matching. For the fuzzy matching the 'Fuzzywuzzy' Python package [10] is used, which uses the Levenshtein string distance to give a percentage of overlap between two strings. This distance measure computes the minimal amount of edits necessary to get from one string to another. Through trial and error a boundary of an 80% or greater overlap has been established. If the two strings get a score of 80% or more it is counted as a match and correctly extracted metadata.
- *Influence of external knowledge*: The extracted metadata is evaluated both with and without the external knowledge. For *author* and *city_of_publication* only the extracted data with a confidence score of either 0, 1, 2 or 'Unknown' are used to evaluate the influence of the external knowledge. These results are compared with the evaluation results when all the confidence scores are considered.
  The influence of the addition of Named Entities for the *pubisher* and *city_of_publication* is evaluated by comparing the results for the data with and without this added information.

The evaluation is done separately for all the different pieces of meta data. No correlations are made between, for example, extracted *authors* and *titles*. All are considered as being independent from one another.

[10] https://github.com/seatgeek/fuzzywuzzy

5.4 Evaluation results

The results of the evaluation for the three different aspects are given below. An error analysis of these results can be found in Section 5.5.

Most of the tables given below contain percentages for different types of matching, which are calculated for the evaluation data. In these tables the highest percentage is given in bold. Some tables give frequencies, as opposed to percentages. Colours are used in the tables to indicate the different types of matching. The colours mean the following:

- *Exact match*: This match type indicates that there is an exact match between the gold metadata and the extracted metadata.
- *Fuzzy match*: This match type indicates that there is an overlap of at least 80% between the gold metadata and the extracted metadata, but there is no exact match.
- *No match*: This match type indicates that the gold metadata and the extracted metadata differ so much that the overlap is less than 80% and no match is made.

5.4.1 Completeness

With the formation of the bibliographical entries not all of them are recognised and formed. The gold evaluation data contains 100 entries for all volumes, but if missing and extra entries are included this number is not the same for the extracted entries. An overview of these numbers can be found in Table 4. In this table the 'T1' behind 1971 and 1980 refers to the RB-REGX technique and 'T2' to the PCFG.

For the evaluation of the extracted bibliographical metadata only the entries that have data in both the gold and the extracted data are used. This means that the completeness aspect of a volume plays part in the evaluation, and missing and extra entries are not evaluated. For this reason the number of extracted entries is calculated as being the number of gold entries minus the missing entries, the extra entries are disregarded for the evaluation.

Ideally, the number of missing and extra entries should be 0 for all the volumes and techniques. Due to incorrectly formed bibliographical entries this is not the reality, this will be explained in more detail in Section 5.5.1.

Volume	Gold entries	Missing	Extra	Extracted entries
1850	100	12	5	88
1882	100	7	1	93
1931	100	31	5	69
1936	100	26	0	74
1961	100	33	1	67
1971 (T1)	100	7	0	93
1971 (T2)	100	27	0	73
1975	100	0	0	100
1979	100	0	0	100
1980 (T1)	100	0	0	100
1980 (T2)	100	14	0	86

Table 4: For each volume and extraction technique the number of entries in the gold evaluation data, the entries in the extracted data, and the number of missing and extra entries.

5.4.2 Exact versus fuzzy matches

The quality of the OCR influence has impact on the results as the data contains OCR errors which lowers the chance of having exact matches with the gold data. For this reason fuzzy matching is used, which means that results with an overlap of 80% or more, but not exact, with the gold metadata are considered correct. The results for the exact, fuzzy and no match match types for all bibliographical metadata are given in Tables 5 and 6 for the RB-REGX technique. For the PCFG the results can be found in Table 7.

With perfect data and a faultless extraction method the results should be 100% exact matches for all the metadata, or for the last three volumes 100% across the exact and fuzzy matches. This is only reality for the *production_number* for the last three volumes, since this number is used as a link between the extracted data and the gold evaluation data this does not come as a surprise. The exact matches and fuzzy matches are both considered to be correct and the 'no match' matches as incorrect. A piece of metadata is considered to be succesfully extracted when the 'no match' percentage is 30% or less. This means that the majority of the metadata is correct.

Looking at Table 5 it is visible that the highest percentages of 'no match' matches occur for the *city_of_publication* and *publisher* metadata. This im-

plies that this metadata is the hardest to extract from a bibliographical entry, looking at the first five types of metadata. Across all the volumes for the metadata 12 out of the 45 percentages of 'no match' matches are above the 30%. Most of those 12 are for the *city_of_publication* and *publisher* metadata, 9 out of 12. The other three are for the *author* and *year_of_publication*.

In Table 6 the highest percentages of 'no match' matches can be found in the column with the *size_of_book* metadata. This performance, however, is explainable and will be described in detail in the error analysis in Section 5.5. Besides the *size_of_book*, the *price_of_book* also contains relatively many 'no match' matches compared to the other metadata.

A big difference can be seen between the performance of the two volumes in Table 7. Where the 1971 volume has very high percentages of 'no match' matches for almost all the metadata, the 1980 volume has low percentages for this matching type and stays below the 30% threshold for all metadata but one, *price_of_book*.

5.4.3 Influence of external knowledge

To evaluate the influence of external knowledge on the extracted metadata, the extracted metadata from the evaluation set has been enriched where the gazetteers have provided confidence scores or the NER has provided suggestions for unfound metadata.

The percentages work the same as mentioned in the previous section, where 30% or less of 'no match' matches is considered a good score. For Tables 8 and 11 the 'T1' refers to the RB-REGX technique and the 'T2' to the PCFG. With the use of the external knowledge the percentages should either remain the same or improve the performance.

5.4.3.1 Gazetteers

The gazetteers are used to enrich the *author* and *city_of_publication* metadata. In Table 8 an overview is shown of the frequency of each confidence score. In this table the confidence score of 0 is the most important, since this score gives an indication of the scope of extracted metadata that is potentially incorrect. This number should be preferably be 0 or close to 0.

In Tables 10 and 9 the 'original' results without confidence scores are given as well as the percentages for the data enriched with the confidence scores based on gazetteers. Table 10 contains the results for the RB-REGX technique and Table 9 those of the PCFG.

Volume	Bibliographical metadata				
	Author	*Title*	*City_of_publication*	*Publisher*	*Year_of_publication*
1850	**0.59**	0.25	**0.48**	0.03	**0.82**
	0.09	**0.59**	0.05	**0.53**	0.00
	0.32	0.16	0.47	0.44	0.18
1882	**0.75**	**0.56**	**0.60**	0.11	**0.65**
	0.11	0.36	0.14	**0.61**	0.00
	0.14	0.08	0.26	0.28	0.35
1931	**0.59**	**0.60**	**0.58**	0.03	**0.80**
	0.14	0.33	0.22	**0.56**	0.00
	0.26	0.07	0.20	0.41	0.20
1936	**0.82**	**0.50**	**0.63**	0.20	**0.85**
	0.06	0.32	0.15	**0.50**	0.00
	0.12	0.18	0.22	0.30	0.15
1961	**0.85**	**0.64**	**0.57**	0.12	**0.72**
	0.12	0.30	0.07	**0.49**	0.00
	0.03	0.06	0.36	0.39	0.28
1971	**0.76**	**0.72**	**0.57**	0.30	**0.78**
	0.10	0.23	0.17	**0.41**	0.00
	0.14	0.05	0.26	0.29	0.22
1975	0.16	0.45	0.25	0.30	**0.75**
	0.73	**0.51**	0.12	0.29	0.00
	0.11	0.04	**0.63**	**0.41**	0.25
1979	0.22	**0.49**	**0.47**	0.34	**0.58**
	0.54	0.39	0.08	0.21	0.00
	0.24	0.12	0.45	**0.45**	0.42
1980	0.21	**0.65**	**0.76**	**0.51**	**0.84**
	0.66	0.29	0.08	0.34	0.00
	0.13	0.06	0.16	0.15	0.16

Table 5: Overview of the percentages of each match type for all the volumes for the following metadata: *author*, *title*, *city_of_publication*, *publisher* and *year_of_publication*

Volume	Bibliographical metadata				
	Size_of_book	*Price_of_book*	*Number_of_pages*	*ISBN_number*	*Production_number*
1850	0.01	**0.48**	n/a	n/a	n/a
	0.00	0.05	n/a	n/a	n/a
	0.99	0.47	n/a	n/a	n/a
1882	0.02	**0.61**	n/a	n/a	n/a
	0.00	0.09	n/a	n/a	n/a
	0.98	0.30	n/a	n/a	n/a
1931	0.26	0.41	**0.67**	n/a	n/a
	0.00	0.06	0.00	n/a	n/a
	0.74	**0.53**	0.33	n/a	n/a
1936	**0.58**	**0.66**	**0.81**	n/a	n/a
	0.19	0.04	0.00	n/a	n/a
	0.23	0.30	0.19	n/a	n/a
1961	**0.72**	**0.76**	**0.78**	n/a	n/a
	0.01	0.15	0.00	n/a	n/a
	0.27	0.09	0.22	n/a	n/a
1971	**0.90**	**0.78**	**0.85**	**0.77**	**0.88**
	0.00	0.00	0.00	0.00	0.00
	0.10	0.22	0.15	0.23	0.12
1975	0.00	0.24	**0.83**	**0.73**	**1.00**
	0.01	0.09	0.01	0.01	0.00
	0.99	**0.67**	0.16	0.26	0.00
1979	**0.49**	0.46	**0.50**	**0.87**	**1.00**
	0.16	0.04	0.06	0.03	0.00
	0.35	**0.50**	0.44	0.10	0.00
1980	**0.81**	0.40	**0.76**	**0.89**	**1.00**
	0.12	0.06	0.07	0.04	0.00
	0.07	**0.54**	0.17	0.07	0.00

Table 6: Overview of the percentages of each match type for all the volumes for the following metadata: *size_of_book*, *price_of_book*, *number_of_pages*, *ISBN_number* and *production_number*

Volume	Bibliographical metadata				
	Author	*Title*	*City_of_publication*	*Publisher*	*Year_of_publication*
1971	0.14	**0.60**	0.11	0.03	0.22
	0.70	0.33	0.03	0.03	0.00
	0.16	0.07	**0.86**	**0.94**	**0.78**
1980	0.21	**0.53**	**0.48**	**0.43**	**0.76**
	0.69	0.47	0.23	0.30	0.00
	0.10	0.00	0.29	0.27	0.24

Volume	Bibliographical metadata				
	Size_of_book	*Price_of_book*	*Number_of_pages*	*ISBN_number*	*Production_number*
1971	0.15	0.33	0.16	**0.58**	**0.68**
	0.00	0.01	0.00	0.00	0.21
	0.85	**0.66**	**0.84**	0.42	0.11
1980	0.08	0.35	**0.67**	**0.74**	**1.00**
	0.64	0.25	0.05	0.03	0.00
	0.28	**0.40**	0.28	0.23	0.00

Table 7: Overview of the percentages of each match type for 1971 and 1980 for all the bibliographical metadata.

	Author				**City_of_publication**			
Volume	*2*	*1*	*0*	*Unknown*	*2*	*1*	*0*	*Unknown*
1850	56	9	0	23	31	21	25	11
1882	55	17	2	19	37	35	16	5
1931	45	19	1	4	41	6	10	12
1936	41	12	0	21	36	12	16	10
1961	59	2	0	6	36	8	17	6
1971 (T1)	62	5	1	25	47	19	23	4
1971 (T2)	49	3	0	21	2	1	1	69
1975	68	8	1	23	65	12	22	1
1979	55	1	0	44	63	0	20	17
1980 (T1)	68	6	0	24	87	1	8	4
1980 (T2)	58	6	0	22	64	2	1	19
Total	616	88	5	232	509	117	159	158

Table 8: Overview of the frequency of each confidence score for *author* and *city_of_publication* for all volumes and different extraction techniques, calculated for the evaluation data without missing entries.

Volume	Without confidence		With confidence	
	Author	*City_of_publication*	*Author*	*City_of_publication*
1971	0.14	0.11	0.14	0.11
	0.70	0.03	**0.70**	0.01
	0.16	**0.86**	0.16	**0.88**
1980	0.21	**0.48**	0.21	**0.48**
	0.69	0.23	**0.69**	0.22
	0.10	0.29	0.10	0.30

Table 9: Percentages per match type for the *author* and *city_of_publication* metadata with and without confidence scores, for the data extracted by the PCFG.

Looking at Table 9 the performance for the *author* metadata stays the same for both volumes with and without confidence scores. For the *city_of_publication* a change in the performance is visible, however, this change is negative since the performance worsens instead of improving.

The opposite of this can be seen in Table 10 where the performance either remains the same or it improves. The improvement is most apparent for the *city_of_publication*.

5.4.3.2 Named Entity Recognition

The Named Entity Recognition is used to enrich the *city_of_publication* and *publisher* metadata. In Table 11 an overview is shown of the frequency of each confidence score.

In Tables 13 and 12 the 'original' results without filled gaps by NER are given as well as the results for the data enriched with the NER with filled gaps for the *city_of_publication* and *publisher*. Table 13 contains the results for the RB-REGX technique and Table 12 those of the PCFG.

Looking at Table 12 the gaps that are filled by NER improve the performance of the *city_of_publication* and the *publisher* for both volumes. The biggest difference is visible for the 1971 volume for the *city_of_publication* which goes down from 86% 'no match' matches to 48%.

This big improvement is not visible in Table 13 where the performance either stays the same or it is worse compared with the scores where no gaps have been filled.

Volume	Without confidence		With confidence	
	Author	*City_of_publication*	*Author*	*City_of_publication*
1850	**0.59**	**0.48**	**0.59**	**0.63**
	0.09	0.05	0.09	0.05
	0.32	0.47	0.32	0.32
1882	**0.75**	**0.60**	**0.77**	**0.73**
	0.11	0.14	0.11	0.08
	0.14	0.26	0.12	0.19
1931	**0.60**	**0.58**	**0.60**	**0.68**
	0.15	0.22	0.15	0.22
	0.25	0.20	0.25	0.10
1936	**0.82**	**0.63**	**0.82**	**0.81**
	0.06	0.15	0.06	0.07
	0.12	0.22	0.12	0.12
1961	**0.85**	**0.57**	**0.85**	**0.76**
	0.12	0.07	0.12	0.06
	0.03	0.36	0.03	0.18
1971	**0.76**	**0.57**	**0.77**	**0.76**
	0.10	0.17	0.10	0.10
	0.14	0.26	0.13	0.14
1975	0.16	0.25	0.16	0.29
	0.73	0.12	**0.73**	0.14
	0.11	**0.63**	0.11	**0.57**
1979	0.22	**0.47**	0.22	**0.58**
	0.54	0.08	**0.54**	0.10
	0.24	0.45	0.24	0.32
1980	0.21	**0.76**	0.21	**0.83**
	0.66	0.08	**0.66**	0.66
	0.13	0.16	0.13	0.11

Table 10: Percentages per match type for the *author* and *city_of_publication* metadata with and without confidence scores, for the data extracted by the RB-REGX technique.

Volume	City_of_publication	Publisher
1850	5	4
1882	2	2
1931	5	1
1936	3	8
1961	2	1
1971 (T1)	2	2
1971 (T2)	58	53
1975	1	1
1979	3	3
1980 (T1)	4	3
1980 (T2)	19	12
Total	104	90

Table 11: Overview of the frequency of the Named Entity Recognition filling an 'Unknown' *city_of_publication* and *publisher* for all volumes and different extraction techniques, calculated for the evaluation data.

	No gaps filled		Gaps filled by NER	
Volume	*City_of_publication*	*Publisher*	*City_of_publication*	*Publisher*
1971	0.11	0.03	0.37	0.06
	0.03	0.03	0.15	0.08
	0.86	**0.94**	**0.48**	**0.86**
1980	**0.48**	**0.43**	**0.56**	**0.47**
	0.23	0.30	0.31	0.31
	0.29	0.27	0.13	0.22

Table 12: Percentages per match type for the *city_of_publication* and *publisher* metadata with and without gaps filled by Named Entity Recognition, for the data extracted by the PCFG.

Volume	No gaps filled		Gaps filled by NER	
	City_of_publication	*Publisher*	*City_of_publication*	*Publisher*
1850	**0.48**	0.03	**0.50**	0.03
	0.05	**0.53**	0.06	**0.52**
	0.47	0.44	0.44	0.45
1882	**0.60**	0.11	**0.61**	0.11
	0.14	**0.61**	0.14	**0.61**
	0.26	0.28	0.25	0.28
1931	**0.58**	0.03	**0.54**	0.03
	0.22	**0.56**	0.20	**0.55**
	0.20	0.41	0.26	0.42
1936	**0.63**	0.20	**0.59**	0.19
	0.15	**0.50**	0.14	**0.54**
	0.22	0.30	0.27	0.27
1961	**0.57**	0.12	**0.55**	0.12
	0.07	**0.49**	0.08	**0.49**
	0.36	0.39	0.37	0.39
1971	**0.57**	0.30	**0.57**	0.30
	0.17	**0.41**	0.17	**0.41**
	0.26	0.29	0.26	0.29
1975	0.25	0.30	0.26	0.30
	0.12	0.29	0.12	0.30
	0.63	**0.41**	**0.62**	**0.40**
1979	**0.47**	0.34	**0.47**	0.34
	0.08	0.21	0.09	0.21
	0.45	**0.45**	0.44	**0.45**
1980	**0.76**	**0.51**	**0.77**	**0.51**
	0.08	0.34	0.09	0.34
	0.16	0.15	0.14	0.15

Table 13: Percentages per match type for the *city_of_publication* and *publisher* metadata with and without gaps filled by Named Entity Recognition, for the data extracted by the RB-REGX technique.

5.5 Error analysis

This section discusses the evaluation results given in Section 5.4. Examples of mistakes are given and described for the three aspects. The frequency of certain mistakes are given as well as the reasoning behind them.

5.5.1 Completeness

With perfectly formed bibliographical entries the number of missing and extra entries should be 0. However, some bibliographical entries are not formed correctly which leads to the missing and extra entries.

For the RB-REGX technique, missing entries are caused by the improper formation of a bibliographical entry, mainly because of incorrectly recognised dashes that have been concatenated instead of forming a new entry. This is visible in Table 4 where the missing entries occur up to 1971, which is the last volume that uses dashes. Of the 33 missing entries for the 1961 volume, 12 of them are caused by the incorrect formation of entries due to dashes of the same author. The same reason holds for 1931 and 1936.

The extra entries are formed by either the incorrect replacement of a dash and/or lines that are in the correct alphabetical order and therefore accidentally taken as a new entry.

The missing entries for the PCFG technique are a sum of the entries that are already missing due to the reasons described above and entries that cannot be parsed by the grammar. This explains why 1980 has 0 missing entries for the RB-REGX technique and 14 for the PCFG.

Examples of missing and extra entries due to the incorrect formation of bibliographical entries can be seen in Figure 15. The missing entry, marked in orange, starts with a dash and is added to the previous item. An extra entry is created because 'Scheveningen' starts with an 'S' and the next entry starts with a dash. Because of this dash the line is not concatenated with the correct entry, the book written by Johannes van der Spek. As a result we get an extra entry and an incorrect *author* for the book 'Het gevoel'.

5.5.2 Exact versus fuzzy matches

The fuzzy matching accounts for the OCR errors that are present in the data. An example of a *title* that contains an OCR error, and is deemed correct by the fuzzy matching, can be seen in (9). It demonstrates that the capital letter 'H' is recognised as a 'B', which is visually similar to an 'H'.

```
Spek (Jae. van der), Bijdra.ge tot de kennisvan de zure gronden in het
Ned. alluvium.Zie Verslagcn van landbouwk. onderzoekingen. 40B. -On
osteogenic sarcoma. (Proefschrift,Universiteit Amsterdam). Utrecht,
Kemink& 'Loon [thans Uitgeverij Broekhoff]. 1933.80. (6 en 219, m.
afb. op 14 pltn.). Niel in den handed.

Spek (Johannes van der), De beteekenis der zielkunde voor de
schoolpraktijk. [2c dr.].

Scheveningen, J. van Bleek (Bootsma & Co.).1931. 80. (48). fl. 1.25.

Scheveningen, Het gevoel. (De gevoels-elementen).In het bijzonder het
religieuze gevoel. Eentweetal aan elkaar aansluitende proefschriften
ter verkrijging van den graad van doctorin do geneeskunde en dien van
doctor in de gudgeleerdheid aan de Rliksuniversiteitto Utrecht.
[Gron., P. Noordhoff]. 1931. Gr.80. (18 on 243 ; 48 en :312, m.
Literatimr'listen en atlas : 8 on 106). 11. 9.-. Zie. Maeder (A.) -
Molenaar (P. J.) -Stunenleving (Onzo) in nood.
```

Figure 15: Entries are missing due to the incorrect formation of previous entries, either 'normal' or because of dashes.

(9) *Gold title*: De verschijning van den **H**eer aan Thomas
Extracted title: De verschijning van den **B**eer aan Thomas

Looking at the evaluation results in Table 5 to 7, it becomes clear that using fuzzy matching is an effective manner to deal with OCR errors. For the RB-REGX technique is does not go below 5% and for the PCFG not below 3%, for the first four pieces of metadata. These four, *author*, *title*, *city_of_publication* and *publisher*, are most likely to contain OCR errors since they can be long and contain letters. The other metadata mainly consists of numbers, is generally shorter and has a fixed structure. This results in them being either an exact match or incorrect.

The *author* metadata shows an interesting trend where the majority of the extracted *authors* is an exact match, with exception of the last three volumes where the highest percentage is for the fuzzy matching. These three volumes are evaluated on the already digitally available gold evaluation data. This trend brings to light that the author in the library catalogue is notated in a different manner than in the Brinkman catalogue. This results in a relatively low exact matching percentage and a relatively high fuzzy matching percentage. An example of this is shown in (10).

(10) *Gold author*: Carl Cornil
Extracted author: Cornil, Carl

Given that the majority of the extracted *authors* is an exact match for the RB-REGX technique, the high fuzzy matching percentage for the *author* metadata extracted by the PCFG stands out. The reason for this is also a different notation of the metadata, see (11). This is caused by the method that extracts the metadata from the parse trees, it inserts a white space between all elements.

(11) *Gold author*: Murdoch, Iris
Extracted author: Murdoch , Iris

5.5.3 Influence of external knowledge

The influence of adding external knowledge is based on the evaluation results by comparing the results without the external knowledge with the results with integrated external knowledge.

5.5.3.1 Gazetteers

The *author* and *city_of_publication* checked against gazetteers and given a confidence score. The frequency of each score is shown in Table 8. Looking at final row of this table it becomes clear that the confidence score of 0 is rare for *authors*, it only occurs 5 times, whereas it occurs 159 times for the *city_of_publication*. The reason for which is that as the first element of the bibliographical entry it is less likely to get an incorrect *author* since it is easier to extract than the *city_of_publication*. Because of this the percentages for *author* with and without confidence scores barely differ, and when they do they slightly improve. An example of an extracted *author* that causes this slight improvement is 'Archives du musée Teyler. Haarl.'. This is not an *author* and it should be 'Unknown', this is caused by the incorrect application of one of the rules.

Looking at Table 10, the *city_of_publication* shows an improvement of at least 4% for the exact matches, with the lowest improvement from 0.25 to 0.29 for 1975 and the highest for both 1961 and 1971 where the percentage goes up from 0.57 to 0.76. This shows that many of the extracted *city_of_publication* are indeed incorrect, when given the confidence score of 0. For example, '91 blz', 'XVIII, 386 blz' and '[Vert' are among the extracted *city_of_publication* values.

The performance of the PCFG extracted metadata goes down when the confidence scores are used, see Table 9. This means that the extracted

city_of_publication that has a confidence score of 0 is actually correct. When looking at these two cases more closely the reason behind this is found. The extracted *city_of_publication* from the 1971 volume is 'Gravenh .', it is given the confidence of 0 because it contains non-alphabetic characters: the white space and the full stop. Similarly, 'The Hague' is given a confidence score of 0, because it is not present in the gazetteer and contains a white space.

5.5.3.2 Named Entity Recognition

The filling of gaps for the *city_of_publication* and *publisher* is relatively infrequent for the RB-REGX technique compared with the PCFG. This can be seen in Table 11, where the highest frequency for the RB-REGX is 8 for the *publisher* of 1936 while for the PCFG it is 58 for the *city_of_publication* of 1971. The metadata is only provided when the value is 'Unknown'. For the metadata extracted with use of the PCFG this occurs often. Especially for the 1971 volume it is frequent, where the biggest improvement can be seen for the *city_of_publication* in Table 12. Examples of filled *city_of_publication* values are 'Amerikaans', 'Baarn', 'Openbare', 'Voorwoord' and 'De Bilt'. Not all filled values are correct but overall it causes an improvement of 38% less 'no match' matches for the 1971 volume, extracted by the PCFG.

For the RB-REGX technique, see Table 13, there is hardly any change in the results when 'Unknown' values for *city_of_publication* and *publisher* are replaced with named entities. For 1936, the exact matches percentage for the *publisher* goes down, when an 'Unknown' is replaced. This indicates that this 'Unknown' value was in fact the correct value. This volume shows a minor improvement for the *publisher*, where the 'no match' percentage drops from 30% to 27% and the fuzzy matches increase from 50% to 54%.

Instances where the percentages drop, for example, the 1931 volume the fuzzy match percentage for *publisher* without NER is 0.56 and with it is 0.55 and for 1850. Another example is the *city_of_publication* for 1961 which drops from 0.57 for exact matches to 0.55. This drop in performance reveals that the extracted 'Unknown' value is the gold evaluation value as well. This is possible since not all metadata is present in all entries, therefore for some metadata 'Unknown' is the correct value.

5.5.4 Reasons for a 'no match'

There are different reasons which result in an incorrect extracted value. Of the in total 2579 'no match' matches between the gold metadata and ex-

tracted metadata the majority of them are caused if either the gold value or the extracted value is 'Unknown' and the other is not. It occurs 396 times that the gold value is 'Unknown' while the extracted value is not and 1333 times the other way around, which accounts for 67% of the 'no match' matches. This reason is the most common one across all the volumes.

For the remaining 'no match' matches a variety of reasons are found. The following reasons are described in more detail, including examples:

- Inadequate gold evaluation data
- OCR errors
- Too few rules
- Wrong application of rule
- Incomplete grammar

5.5.4.1 Inadequate gold evaluation data

This reason only holds for the digitally available evaluation data, since the manually created evaluation data contains all the needed information. The digitally available gold evaluation data does not contain all information for every entry, or it is notated in a different manner. This is the cause of the poor performance for *size_of_book* for 1975. For this volume the Brinkman catalogue notates the size as follows: '21 x 13', while in the gold evaluation data it is notated as '21 cm'. The *price_of_book* is not always present in the gold evaluation data, while it is correctly extracted from the entries. This gives the wrong impression, namely, that the extracted metadata is incorrect. However, the extracted metadata adds to the currently available data.

5.5.4.2 OCR errors

Some metadata is seen as incorrect because it cannot be matched with the gold data due to OCR errors. Examples of this can be seen in (12) and (13), where the OCR recognition is so poor that it is deemed incorrect.

(12) *Gold author*: Holtius (A. C.)
Extracted author: H01tiIIS (A. C.)

In the example in (12) some letters are recognised as numbers, and the lower case 'u' is recognised as two capital 'I' letters. The reverse is true in the example in (13) where the capital 'D' is recognised as a lower case 'b'. Also it is visible that the 'e' and 'o' are similar since it both occurs that an 'e' is recognised as an 'o' and the other way around.

(13) *Gold title*: De roode wagen
Extracted title: be roodo wagon

5.5.4.3 Too few rules

This reason is visible for the *size_of_book* for the three oldest volumes, 1850, 1882 and 1931. The performance for this metadata is very poor, the ones that are correct are caused by 'Unknown' being correct for a few entries. The size indication in these older volume differs a lot from the other volumes, '8o' is a size indication as opposed to '21 x 16' or '23 cm' in the more recent versions.

5.5.4.4 Wrong application of rule

It occurs that the rule is applied to an incorrect part of the entry. Two examples of *authors* can be seen below in (14) and (15), which should not have been recognised as such.

(14) *Gold author:* Unknown
Extracted author: Archives du musée Teyler. Haarl.

(15) *Gold author*: Unknown
Extracted author: Vertellingen voor het kleine christenvolkje. Gorinch.

These kind of mistakes should be avoided, to accomplish this a maximum length for an *author* is set, unfortunately in these examples the separating character normally between the *title* and *author* falls within this limit without it being an *author* and a *title*. This does not only have consequences for the *author*, in these cases this mistake also causes the *title* and *city_of_publication* to be incorrect.

5.5.4.5 Incomplete grammar

The grammars created can not parse all the entries. This is due to missing production rules, an entry can only be parsed if its composition matches the production rule exactly. To try to minimise the impact of this a rule is added that only extracts the *author*, *title* and *production_number*, this leaves a lot of metadata 'Unknown'.

6 Discussion

This section discusses some general results, the general performance of the text mining techniques and the best evaluation method.

6.1 General results

Some general results per volume are given below, including the number of extracted bibliographical entries and the average per year for each volume (Table 14) and the number of entries from which a particular piece of metadata has been extracted (Table 15 and 16), calculated for the rule-based and regular expressions technique.

Volume	Total entries	Average per year
1850	40301	1221
1882	14819	1482
1931	22707	4541
1936	23202	4640
1961	50299	10060
1971	12707	12707
1975	15219	15219
1979	22170	22170
1980	16930	16930

Table 14: For each volume the total number of extracted entries and the average number of entries per year per volume.

In table 14 the total number of extracted bibliographical entries from each volume is given, alongside an average of entries per year for the particular volume. A big variance is visible in the total number of entries per volume, since not all volumes cover an equal amount of years an average per year gives a better insight. Looking at these numbers there is a clear increase in the number of entries per year per volume. An exception to this trend is the 1980 volume, however, this volume is incomplete and the OCR output stops at the letter 'P' in the alphabet. If it were complete it would exceed the number of entries from 1979. This trend is as can be expected given the progress and innovation in book printing and publishing.

	Extracted bibliographical metadata				
Volume	*Author*	*Title*	*City_of_publication*	*Publisher*	*Year_of_publication*
1850	0.80	1.00	0.88	0.88	0.83
1882	0.85	1.00	0.94	0.94	0.83
1931	0.91	1.00	0.84	0.84	0.70
1936	0.76	1.00	0.88	0.88	0.77
1961	0.68	1.00	0.88	0.88	0.67
1971	0.71	1.00	0.93	0.93	0.69
1975	0.68	1.00	0.95	0.95	0.69
1979	0.45	0.79	0.81	0.81	0.54
1980	0.56	0.79	0.94	0.94	0.67

Table 15: Overview of the percentages of extracted metadata from entries for all the volumes for the following metadata: *author*, *title*, *city_of_publication*, *publisher* and *year_of_publication*.

	Extracted bibliographical metadata				
Volume	*Size_of_book*	*Price_of_book*	*Number_of_pages*	*ISBN_number*	*Publication_number*
1850	0.00	0.50	n/a	n/a	n/a
1882	0.00	0.85	n/a	n/a	n/a
1931	0.00	0.62	0.55	n/a	n/a
1936	0.72	0.66	0.55	n/a	n/a
1961	0.76	0.64	0.65	n/a	n/a
1971	0.79	0.65	0.80	0.27	0.89
1975	0.72	0.66	0.81	0.32	0.75
1979	0.59	0.63	0.59	0.54	0.83
1980	0.83	0.62	0.83	0.52	0.85

Table 16: Overview of the percentages of extracted metadata from entries for all the volumes for the following metadata: *size_of_book*, *price_of_book*, *number_of_pages*, *ISBN_number* and *production_number*.

Taking the evaluation results considering the completeness into account the number of entries will differ from the actual number of entries in the volumes. Entries are missing due to illformed entries, which will mainly be present in the volumes up to 1971.

The percentages in Tables 15 and 16 give an overview of the amount of entries a certain piece of metadata is extracted from. This is calculated by taking the percentage of entries where the metadata is not 'Unknown'. Across the different metadata these percentages vary between 0.27 and 1.00, the lowest percentage being for the *ISBN_number* and the highest for the *title*. It should be mentioned that the goal is not to have a percentage of 1.00 for all the metadata, not all entries contain all the information and therefore it will not be 100% of the entries that have a particular piece of metadata.

An interesting trend is visible for the *city_of_publication* and the *publisher*, all the percentages are equal for the two pieces of metadata across all the volumes. This indicates that the metadata is always extracted together, it is not possible to extract only a *city_of_publication* and not a *publisher* and vice versa.

For almost all the volumes the percentage of the *title* is 1.00, only for 1979 and 1980 it is 0.79. The cause lies with the character that separates the *title* within the bibliographical entry. For the first seven volumes this is a full stop, every entry contains a full stop somewhere which means that a *title* can always be extracted. The final two need to contain a backslash character as a separation for the *title*, when this is not present no *title* will be extracted.

The *size_of_book* has a percentage of 0.00 for the first three volumes, the reason behind this is explained in Section 5.5.4.3.

6.2 Text mining techniques

In Section 4.4 three different techniques are described that are used to extract the metadata from the bibliographical entries. The technique with Named Entity Recognition (4.4.3), is only used to extract the *publisher* metadata and proved to be inconsistent. For this reason only the first two techniques were evaluated. Those two techniques are compared below based on the evaluation results and the implementation.

6.2.1 RB-REGX

This technique is consistent and able to extract metadata from all the entries. Looking at the evaluation results in Tables 5 and 6 for the 1971 and 1980 volumes the average percentage 'no match' matches across all the metadata are 18% and 15% respectively. The results are influenced in a positive way by the use of external knowledge, where the percentage either goes up or remains the same for the gazetteers. The NER is used infrequently and does not influence the results in a major way.

6.2.2 PCFG

The PCFG is not able to extract metadata from all the entries and skips some entries because they cannot be parsed by the grammar. The average percentage of 'no match' matches in Table 7 is 57% for 1971 and 21% for 1980. The gazetteers do not improve the performance as it either remains the same or gets lower. NER is used frequently to replace 'Unknown' values with named entities for the *city_of_publication* and the *publisher*.

Looking at the performance of the two techniques, the RB-REGX technique performs either better or the same as the PCFG. The average percentage of 'no match' matches is lower for the RB-REGX and even when using external knowledge the performance of the PCFG is lower or equal to that of the RB-REGX technique. Furthermore the RB-REGX requires less work to create an apply.

6.3 Evaluation methods

Two types of evaluation methods are used to evaluate the extracted data. One using manually annotated data as gold evaluation data, the other using data from the library catalogue which is already digitally available.

The main difference between the two gold evaluation sets concerns the precision with which it matches with the extracted data. The manually created data is an exact match with the extracted data, without taking OCR errors into account. For the digitally available evaluation data this is not true, the library catalogue uses a different notation for certain metadata compared to the Brinkman catalogue. For example in the Brinkman catalogue 'Amsterdam' is regularly abbreviated to 'Amst', this difference in

notation is not picked up by the fuzzy matching. With the manually created evaluation data this does not cause a problem since the evaluation data has 'Amst', however, the digitally available data does not and will return this as incorrectly extracted metadata.

Another disadvantage of the digitally available evaluation data is that it is not necessarily in the same order as it is in the catalogue. This results in being unable to check the extracted data for any missing or extra entries.

Additionally, the method of manually annotation entries as gold evaluation data is used in the literature about digitising dictionaries, see Section 2.1.4. Even though this is more labour than using already digitally available, if this exists, the manually created evaluation data is of high quality and with certainty matching with what the system should have extracted.

6.4 Usage of extracted metadata

With the extracted metadata converted to the PICA+ format it can easily be compared with the data that is already present in the library catalogue of the National Library of the Netherlands. With this comparison the library catalogue can be expanded by adding, for example, the *price_of_book* for publications for which this metadata is extracted but not yet present in the library catalogue. Additionally, publications can be identified that are not yet in the collection of the National Library and steps can be taken to obtain them and add them to their collection.

The most important metadata for to achieve this are the *author* and *title* of a publication, also the *year_of_publication* to account for multiple prints of the same publication. The OCR errors will influence this process but the fuzzy matching helps with this. The metadata that is correct with fuzzy matching can be used since it is correct and the correct metadata can most of the time easily be spotted by a human. Since all the entering of publications into the library catalogue is done manually this should pose no problem.

7 Conclusion

This research explored different text mining techniques that can be used to extract data from a digitised catalogue and to structure the data. The research question that guided this research reads:

Which text mining techniques can be used to structure digitised bibliographical data and what is the best way to evaluate these methods?

The biggest problem when working with the OCR output was creating the bibliographical entries. This has been accomplished by combining manual annotation and automatic concatenation of lines based on the alphabetic nature of the catalogue. Another problem is caused by the OCR errors within the data, this is largely resolved by allowing fuzzy matching during the evaluation process.

Three text mining techniques have been used to extract metadata from bibliographical entries: a Rule-based system with Regular Expressions, a Probabilistic Context-Free Grammar and Named Entity Recognition. From the three text mining techniques the Rule-based with Regular Expressions technique is the most consistent and efficient one. In combination with the external knowledge this technique has the best results and the method can easily be adapted and applied to other volumes and other catalogues. The best method to evaluate the extracted data is by manually creating the evaluation data form the original catalogue. This ensures that the correct notation is used and can identify missing entries and incorrectly formed additional entries.

In future work this research can be expanded to include other types of catalogues such as the Brinkman topic catalogues or auction catalogues. The current results can be improved by using data with fewer OCR errors, especially for the older volumes, and include the volumes that are now lost due to their poor OCR quality. Additionally, with annotated data a machine learning system can be developed to form the entries and extract the data from them.

References

Bago, P., & Ljubešić, N. (2015). Using machine learning for language and structure annotation in an 18th century dictionary. *Electronic lexicography in the 21st century: linking lexical data in the digital age*, 427–442.

Breuel, T. M. (2007). The hocr microformat for ocr workflow and results. In *Ninth international conference on document analysis and recognition (icdar 2007)* (Vol. 2, pp. 1063–1067).

Karagol-Ayan, B., Doermann, D., & Dorr, B. J. (2003). Acquisition of bilingual mt lexicons from ocred dictionaries. In *Proceedings of the 9th mt summit* (pp. 208–215).

Khemakhem, M., Foppiano, L., & Romary, L. (2017). Automatic extraction of tei structures in digitized lexical resources using conditional random fields. In *electronic lexicography, elex 2017.*

Ma, H., Karagol-Ayan, B., Doermann, D., Oard, D., & Wang, J. (2003). Parsing and tagging of bilingual dictionaries. *Traitement Automatique Des Languages*, *44*(2), 125–149.

Manning, C. D., & Schütze, H. (1999). *Foundations of statistical natural language processing*. MIT press.

Maxwell, M., & Bills, A. (2017). Endangered data for endangered languages: Digitizing print dictionaries. In *Proceedings of the 2nd workshop on the use of computational methods in the study of endangered languages* (pp. 85–91).

Nadeau, D., & Sekine, S. (2007). A survey of named entity recognition and classification. *Lingvisticae Investigationes*, *30*(1), 3–26.

Romary, L., & Lopez, P. (2015). Grobid-information extraction from scientific publications.

Thompson, K. (1968). Programming techniques: Regular expression search algorithm. *Communications of the ACM*, *11*(6), 419–422.

Traub, M. C., Van Ossenbruggen, J., & Hardman, L. (2015). Impact analysis of ocr quality on research tasks in digital archives. In *International conference on theory and practice of digital libraries* (pp. 252–263).

Veen, H. v. d., & Waterschoot, J. v. (2001). *De brinkman: ondankbare arbeid voor nu en voor later*. Alphen aan den Rijn: Kluwer.

A PICA+ output

004A $ffl. 2.25
011@ $a1937
021A $a@Vlucht met Elisabeth
028A $dW.$aAckermann
033A $pAmst$nAndries Blitz
034D $a211
034I $20 x 125

004A 09026807708ffl. 24
006C $0B7503565
011@ $a1974
021A $a@Afstand van vermogensrechten
028A $dHendrik Antonius Maria$aAaftink
033A $pDeventer$nKluwer
034D $a141
034I $24 x 16

004A $ff 0.20
011@ $a1864
021A $aDe @verschijning van den Beer aan Thomas
028A $dL. G.$aJames

004A $ffl. 6.90
006C $0B7153864
011@ $a1971
021A $aDe @zwarte mus
028A $dSibe$cvan$aAangium
033A $p's-Gravenh$nJ
034D $a172
034I $22 x 15

B Deliverables

The deliverables of this research can be found on https://github.com/Karen-GH/Master-Thesis. The deliverables are both code and data/output.

Code:

- *Form_bibliographical_entries.py*: Forming bibliographical entries
- *Reform_bibliographical_entries.py*: Reforming bibliographical entries that are not alphabetical
- *Replace_dashes.py*: Replacing the dashing with correct information
- *Extract_metadata_RB-REGX.py*: Extract metadata using the RB-REGX technique
- *Convert_PICA.py*: Convert to PICA+ format

Data:

- *brin003197101.pdf*: Orignal PDF of catalogue volume
- *brin003197101.txt*: Original OCR output of catalogue volume
- *197101_letter_sections.txt*: Processed OCR with letter sections
- *Formed_bib_entries_197101.txt*: Bibliographical entries formed using the first letter
- *Reformed_bib_entries_197101.txt*: Bibliographical entries reformed to be alphabetical
- *Replaced_dashes_197101.txt*: Bibliographical entries with replaced dashes
- *Extracted_metadata_1971.tsv*: Extracted metadata using the RB-REGX technique
- *Extraced_metatdata_1971_PICA.txt*: Extracted metadata in PICA+ format