Opleiding Informatica

Analysis of extracted references to case law in Dutch court decisions

Jonathan Strijker

Supervisors:
Suzan Verberne (LIACS, Leiden University)
Gijs van Dijck (Maastricht Law & Tech Lab, Maastricht University)

BACHELOR THESIS

Leiden Institute of Advanced Computer Science (LIACS)
www.liacs.leidenuniv.nl   01/07/2021

Abstract

In the Dutch legal system, a judge may cite a previous judgment in his judgment for a great number of reasons. The citing of these previous judgments, sometimes referred to as jurisprudence or case law, is a common occurrence in judicial decisions. These judicial decisions are available in digital form on the website '`Rechtspraak.nl`'. A software framework to recognize and resolve the references to case law in these digital judicial decisions exists. The fruits of this framework can be seen on the website '`Linkeddata.overheid.nl`'. On here, one can view the processed judicial decisions along with the detected references marked in the text. A dump of the database behind the site has been made available to the public. The primary goal of this thesis is to analyse the extracted references in that database. Therefore, I will first go over the framework itself and how it works. Subsequently, I will describe the characteristics of the judicial decisions that we are working with. Furthermore, I will be explaining the methods that I have used for the preprocessing, extraction and analysis of the data. Finally, I will extensively describe all types of errors I came across. I will be giving special attention to the type of error that occurs when a text is falsely detected as a reference; the false positives.

Contents

1 Introduction

In this thesis, our primary objects of interest are Dutch court decisions. A court decision is a decision, sometimes referred to as a judgment or verdict, rendered by a judge in civil, administrative or criminal proceedings. All court decisions are recorded in writing [1] and the collection of all court decisions is referred to as jurisprudence [2], or 'case law' . Within the judicial system of the Netherlands, there are several legal grounds for the publication of these court decisions including: Article 6 of the European Convention on Human Rights (ECHR), Article 121 of the Constitution of the Netherlands and art. 14(1) of the International Covenant on Civil and Political Rights (ICCPR) [3]. There has been a lot of debate around the scope and reach of these articles but their main goal is clear: to protect citizens from secret trials[1] and to provide an opportunity for the judiciary[2] to be subject of public scrutiny [4].

Court decisions are published in several law reviews[3], case law bundles and databases [5]. A distinction can be made between publication by commercial publishers and publication by the government. In the past, publication of court decisions was primarily done by commercial publishers. However, since the arrival of "Rechtspraak.nl" in 1999, that is now no longer the case [6]. Rechtspraak.nl is the official website of the Judiciary of the Netherlands and it contains a database with publicly available court decisions [7].

There are several national legal systems in the world and generally a distinction can be made between two types of systems: common law systems and civil law systems. The Dutch legal system is based on the civil law system; this means that originally the system was built on codified statutes as opposed to being built on previous court decisions. Most countries, including the Netherlands, use a mix of features from both law systems [8]. Therefore, jurisprudence, the collection of all judgments, is seen as another source of law in the Netherlands [9].

In the Netherlands, judges are tasked with applying the law. They should base their judgments on the law of the Netherlands and therefore, they may choose to use previous judgments in their verdicts. However, as opposed to common-law systems, judges in the Netherlands are not 'bound' by law to follow previous judgments from higher courts [9]. They can use them as guidelines, but they may also choose to diverge from them [10]. In both cases the judge can use a previous judgment by citing it in his judgment. This citing happens on a regular basis and it is done for many different reasons [11].

The analyses of these 'citations', or references to previous cases in court decisions was in my eyes an area in which much was still unexplored and therefore, I believed it to be filled with many research opportunities. However, as I will explain throughout this paper, analysing these type of references is not as straightforward as one might expect. One would have to first detect and extract these references in the court decisions. We will discover that this process is not always without errors.

---

[1] "A secret trial is a trial that is not open to the public, nor generally reported in the news, especially any in-trial proceedings." – https://en.wikipedia.org/wiki/Secret_trial

[2] "The judiciary is the system of courts that interprets, defends, and applies the law in the name of the state." – https://en.wikipedia.org/wiki/Judiciary

[3] "A law review (or law journal) is a scholarly journal or publication that focuses on a wide array of legal issues." – https://en.wikipedia.org/wiki/Law_review

1.1 Available data

The data I will work with are Dutch court decisions. However, not all available court decisions are published in the Netherlands. On the 30th of May this year (2021), in an interview with the Dutch newspaper NRC, the chairman of the Council for the Judiciary[4] has stated that less than 5% of all Dutch court decisions get published. This rate can be seen by some as low however, additionally he rightfully states that not all court decisions are interesting enough to publish [12].

A typical example of 'uninteresting' court decisions are "trials in absentia", also known as judgments by default. Usually in trials, a defendant is summoned to court to defend himself against certain claims. If the defendant does not show up, the judge usually accedes the claims against them. And so, in these situations, a judge merely has to verify that the plaintiff's claims are neither unfounded nor unlawful [13]. It should come as no surprise that these type of trials are deemed "almost never" interesting for publication by the Council of the Judiciary [14]. Furthermore, some cases can not be published because of privacy reasons [12].

If we take a closer look at the publication rate for relevant[5] court decisions in the annual report of the Judiciary, we see that in 2020 around the 72 out of every 1000 relevant court decisions get published, see figure 1.[6] This means that a total of $38,000/(72/1000) \approx 528,000$ court decisions from 2020 are seen as relevant by the judiciary. For comparison, there were a total of 1.39 million cases in 2020 [15].

Tabel 19: Totaal van gepubliceerde uitspraken inclusief VK, afgerond op honderd	2007	2016	2017	2018	2019	2020
Aantal gepubliceerde uitspraken	17.100	25.500	27.900	29.900	31.100	38.000
Per 1.000 relevante uitspraken	19	35	41	46	47	72

Figure 1: Publication rate of 'relevant' cases per year [15]

The main reason for this low publication rate is quite simple: all published court decisions need to be anonymized before being published, and the judiciary simply does not have the resources to anonymize all court decisions [16]. This is due to the fact that anonymization has to happen in careful accordance with the 'anonymization guidelines'[7] and that is for now a process primarily done by hand [3].

We can establish that the amount of court decisions we have at our disposal is significantly lower than the amount that we in theory could be using. The low publication rate of court decisions therefore directly affects the amount of data that we can work with in this thesis. This might be something to keep in mind while interpreting the results of this paper.

---

[4] "The Council also acts as a spokesperson for the judiciary on both national and international levels." – https://www.rechtspraak.nl/English

[5] "Disposals of which publication is almost never interesting (judgments by default, withdrawals and disposals without verdict) are left out."

[6] Court decisions by the VK="Vreemdelingenkamer"/Foreign Department are included

[7] https://www.rechtspraak.nl/Uitspraken/paginas/anonimiseringsrichtlijnen.aspx

1.2 Research goal

The research goal of this thesis is to analyse extracted references to case law in Dutch court decisions. As a good method (section 2.1) to extract case law references in court decisions has been developed, I will look at the possibilities of reducing the amount of false positive references that were extracted using that method.

1.3 Contributions

This research will deliver the following items:

1. Code to handle the preprocessing of the available data.

2. Code to visualize the characteristics of the preprocessed data.

3. Code to extract references from the data into CSV files.

4. Code to output irregularities in these extracted references.

5. List of errors discovered in the data.

6. Evaluation of some extracted references.

The code (1-4) can be found on github[8] and the rest (5-6) can be found within this thesis (see section 5).

1.4 Thesis overview

This chapter contains the introduction; Section 2 discusses extra background information and related work; Section 3 describes the data; Section 4 discusses the methods used on the data; Section 5 describes the results; Section 6 discusses the results; Section 7 concludes and considers opportunities for further research.

---

[8]https://github.com/Strijkerr/BachelorThesis

2 Background and related work

In this section we will discuss the most advanced method I have seen so far for extracting references to case law in Dutch court decisions. It is a framework that has been developed in the span of multiple years. We take a look at its components and how it finds references. We will then discuss the type of errors that can occur in the process of finding these references. Furthermore, we will also have a look at a paper that has analysed references to case law in Dutch court decisions before.

2.1 eXtendable Legal Link eXtractor

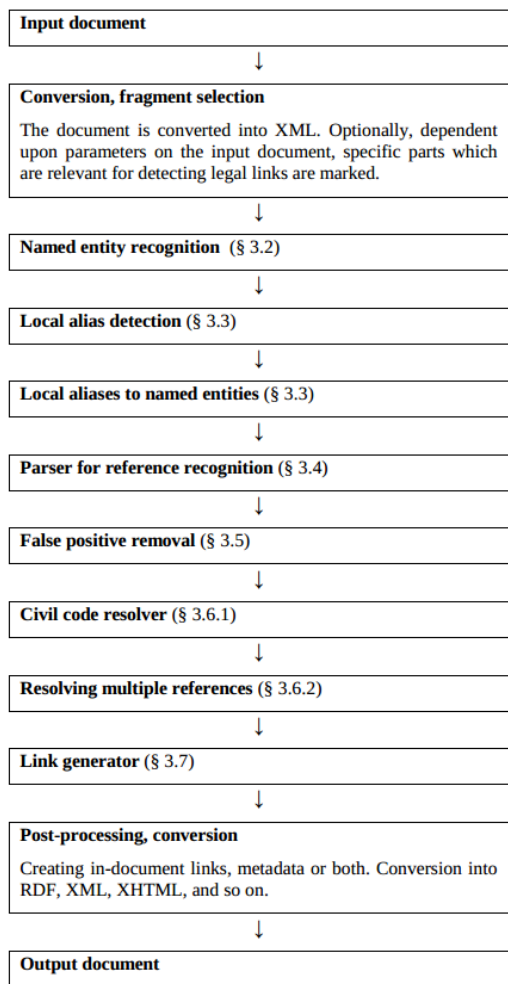| Input document |
| --- |
| ↓ |
| **Conversion, fragment selection**
 The document is converted into XML. Optionally, dependent upon parameters on the input document, specific parts which are relevant for detecting legal links are marked. |
| ↓ |
| **Named entity recognition** (§ 3.2) |
| ↓ |
| **Local alias detection** (§ 3.3) |
| ↓ |
| **Local aliases to named entities** (§ 3.3) |
| ↓ |
| **Parser for reference recognition** (§ 3.4) |
| ↓ |
| **False positive removal** (§ 3.5) |
| ↓ |
| **Civil code resolver** (§ 3.6.1) |
| ↓ |
| **Resolving multiple references** (§ 3.6.2) |
| ↓ |
| **Link generator** (§ 3.7) |
| ↓ |
| **Post-processing, conversion**
 Creating in-document links, metadata or both. Conversion into RDF, XML, XHTML, and so on. |
| ↓ |
| **Output document** |

Figure 2: "Schematic display of pipeline for detecting and resolving references to national legislation." – Opijnen, 2015 [17]

As mentioned in section 1, a tool to extract references to case law in court decisions has already been developed. Van Opijnen has described this software framework in his paper: "Beyond the Experiment: the eXtendable Legal Link eXtractor" that he published in 2015 [17]. This framework is used for "detecting and resolving references to (national and EU) legislation, case law, parliamentary documents and official gazettes." For this thesis, we are investigating its ability to detect and resolve references to case law.

Van Opijnen starts his paper by stating that "the growing public availability of legal documents is a positive development". He then explains that the interlinking of these legal documents is important and that "making explicit all textual links that are not computer readable, would give real added value".

The eXtendable Legal Link eXtractor (xLLx), the topic of the paper, was designed to help in these matters. A display of the pipeline for detecting and resolving references to national legislation can be seen in figure 2. Although the pipeline in the figure is used for a very specific purpose, Opijnen states that many of its main components are used in other pipelines as well: e.g., the pipeline for recognizing references to case law.

We see that the pipeline has a 'named entity recognition' (NER) component that is used for recognizing named entities. In this component a lookup is done to see if a word is in the list of known legal resources. Van Opijnen used a 'trie data structure' for its implementation. He settled on this ordered data structure, a type of search tree, because its capabilities made it well suited for this job. When a text containing a reference to a judicial decision is parsed (e.g. 'Rensing/Polak II'), the trie should detect it as a named entity.

An 'exception list' exists to prevent entities, that appear as common words (e.g., decision), from being included in the trie dataset, as this would lead to many false positives.

'Rensing/Polak II' is the alias of a well-known judicial decision. xLLx can recognize these 'global aliases' as named entities. Besides global aliases, 'local aliases'[9] are also detected by xLLx.

Legal references not using the official or non-official title of a legal document can be recognized using the 'parser for reference recognition' component. It uses pattern recognition for recognizing these type of legal references. Regular expressions[10] have been tested for use in this component by Van Opijnen. However, he noticed that they had serious drawbacks when it came to "recognizing legal references in a large-scale environment". Therefore, he chose to use 'parsing expression grammar'[11] for this component instead.

The next component in the pipeline is the 'false positive removal' component. Van Opijnen states that "false positives occur" and that "without context analysis, the difference is hard to tell". The goal of this component is pretty self-explanatory. More on this component can be seen in section 2.2.

A component that is not found in figure 2, but that Van Opijnen does discuss in his paper, is a specific component used for resolving references to case law. Because case law references in the Netherlands can be very complex, a 'canonicalization' process to detect and resolve those references is needed. Van Opijnen has extensively described this process in a different paper [18].

At the end of the paper, we see that Van Opijnen has measured the recall and precision of xLLx on a sample of "two randomly selected judicial decisions from each of the seven types of courts in the Netherlands". His results can be seen in the table below [17]:

Table 1. Recall and precision of xLLx.

Reference to:	Relevant	Detected	False positive	Recall	Precision
National legislation	271	265	10	98%	96%
EU acts	69	64	1	93%	98%
Elements of (national and EU) legislation	204	183	0	90%	100%
Case law	151	140	1	93%	99%
Parliamentary documents	15	13	0	87%	100%
Official journals	9	9	0	100%	100%
Total	719	674	12	94%	98%

We see that the recall and precision of xLLx for recognizing references to case law are quite impressive. For future work, it might be interesting to test these evaluation metrics of xLLx on some bigger samples.

However, we do not have access to the xLLx software, but we do have access to the database containing all case law documents processed by xLLx called "Linkeddata.overheid.nl" (LiDO).

On that site, one can view case law documents with the detected references, hyperlinked,[12] in the text. Section 3.7 of this thesis is dedicated to describing LiDO further.

---

[9] e.g., "the European Convention on Human Right and Fundamental Freedoms (**hereinafter 'the Convention'**)."

[10] "A regular expression is a sequence of characters that specifies a search pattern." – https://en.wikipedia.org/wiki/Regular_expression

[11] "A parsing expression grammar (PEG) is a formalized machine-oriented syntax introduced by Bryan Ford in 2004" – Opijnen, 2015

[12] "In computing, a hyperlink, or simply a link, is a reference to data that the user can follow by clicking or tapping. A hyperlink points to a whole document or to a specific element within a document." – https://en.wikipedia.org/wiki/Hyperlink

2.2 Types of errors in xLLx

In this next paper by Van Opijnen [19], published 3 years later, he reflects on the xLLx, LiDO and the future of legal information retrieval. Using less technical terms than in the previously discussed paper, he outlines the functioning of the xLLx and its components again. In this paper however, he goes deeper into the special routines that are implemented to tackle false positives in the xLLx.

One of the routines that is implemented to tackle false positives in the recognition of references to court decisions, can best be explained with the help of the following example he gives. 'Cilfit' and 'United Brands' are both aliases of well-known judicial decisions. However, when those aliases appear in texts; "Cilfit" usually is a reference to the judicial decision while that is not always the case for "United Brands". When "United Brands" appears in texts, it might be a reference to the name of the company. He tackles this problem regarding aliases using the following routine:

> "As a solution for optimal balance between false positives and false negatives, the LinkeXtractor can make a distinction between global aliases that have to appear in combination with words like "decision", "case", "verdict", and global aliases that can appear without such additional terms. In which category a specific alias belongs still has to be done by a human editor."

The next step involves verification of found references. Van Opijnen states that references found solely by named entity, do not have to be verified. But if a reference has been recognized through pattern matching, the existence of the recognized resource has to checked.

All of this sadly does not prevent the xLLx from making a few errors. Luckily, Van Opijnen gives a clear overview of the errors one can expect:

- False negatives

- False positives

- Mapping errors

- Ambiguities

- Incomplete or erroneous references

Opijnen states that the 'false negatives', 'false positives' and 'mapping errors' in this list can only be detected by humans. However, I believe it is worth investigating ways to detect these errors automatically. And therefore, finding a way to detect these errors automatically, is one of the goals of this thesis.

By now, we have got a better understanding of the software used for recognizing references to case law in court decisions and additionally, we also have got an idea of the errors it might produce.

2.3 Most commonly referenced court decisions – 2011

Relevant to our work is a paper by Winkels et al. [20], in which Winkels tries to "decide upon the importance of cases". He does this by analysing the network of citations between cases. In one of the two studies he conducted for his paper, he uses 89,179 cases from Rechtspraak.nl ranging from 1999 until 2008. He identifies the references in these cases with a modified version of a priorly developed reference parser for Dutch legislation [21]. Because of certain issues, he ended up using only the LJN and NJ references (see section 3.6) in these cases. He additionally states that at the moment of writing the paper (2011), these reference variants were "the most widely used ones anyways".

Nr	LJN/NJ	Court	Name	PageRank
1	AA 7309/ NJ 2000/721	Supreme Court	Reasonable period	1
2	AG 4158/ NJ 1981/635	Supreme Court	Haviltex	3
3	AU 9130/ NJ 2006/393	Supreme Court	Pot nursery	5
4	AM 2533/ NJ 2004/376	Supreme Court	Waste pipe	9
5	AQ 4473/ NJ 2005/149	CoJ EU	-	2
6	AD 5163/ NJ 2002/317	Supreme Court	-	10
7	AO 1427/ NJ 2005/493	Supreme Court	Pension fund DSM/Fox	6
8	ZC 2536/ NJ 1999/285	Supreme Court	Brok/Huberts	4
9	ZD 0328/ NJ 1996/249	Supreme Court	Zwolsman	-
10	AD 4922/ NJ 2002/82	Supreme Court	Schwarz/Gnjatovic	-

Figure 3: "Top-10 cases 2nd network based on incoming references" – Winkels, 2011 [20]

If we look at figure 3, we see the top 10 most commonly referenced decisions in his study. In this table, we see mostly decisions by the Supreme Court. He states that "all most-cited cases turn out to be from the Supreme Court and the European Court of Justice". He also states that "a sudden drop in the number of citations of a case may be an indication of codification".

It is interesting to see that analysis of references to case law has been extensively described in a paper before. Even though Winkels' goal is different from mine, several similarities exist between his paper and this thesis.

In this thesis I will be working with the court decisions from Rechtspraak.nl as well. However, it must be noted that the collection of decisions I am using, is much more recent. This means that I have access to court decisions from 2008 onwards as well. Seeing as the publication rate of cases has climbed over the last few years (see section 1.1), it probably means that the amount of court decisions I can analyse is much greater ($\geq$ linear increase).

Another difference is that Winkels only analyses LJN and NJ references in his study while I want to analyse all types of references to case law in court decisions. The fact that LJN and NJ references were the most widely used back then, does not guarantee that they still are.

Therefore, taking into account the differences and the fact that some decisions have been codified in the last few years, we might find something interesting in comparing the top 10 most commonly referenced decisions 10 years ago (2011) with the most commonly referenced decisions today (2021).

We have now set the basis for this thesis and we can now continue with the next section; the data.

3 Data

Dutch court decisions can be found in several places (see section 1). One of the easiest ways for the general public to access these court decisions is through the website: "Rechtspraak.nl". Therefore, it made sense to pick Rechtspraak.nl as the source of our court decisions. In this chapter I describe the characteristics of the available court decisions, the format and structure of those decisions and the references we want to extract from them.

3.1 Selection criteria

In section 1.1 we have seen that not all Dutch court decisions are made public because of a lack of resources. Therefore, the judiciary has to make decisions about which cases to publish and which not. The criteria for this selection can be found in a resolution[13] on Rechtspraak.nl. The aim of this resolution is to, in principle, publish all court decisions by the highest courts and to publish the potentially important cases in lower courts [6]. Publication is mandatory for court decisions that satisfy certain specific criteria listed in that resolution.[14]

Does this mean that all court decisions that do not meet these criteria go unpublished and unknown to the public? No, not necessarily. For some judgements, key data like (ECLI (section 3.3), court, case number, date of judgement) and, possible source are published [22]. However, those cases are not interesting to us because we require decisions to have textual parts in them for us to analyse references in. We will have to take this into account when preparing the dataset of court decisions for analysis.

3.2 "Open Data van de Rechtspraak" dataset

There are multiple ways of accessing the court decisions on Rechtspraak.nl. One way is to use the webservice made available by Rechtspraak.nl. However, one could also download the entirety of the available court decisions in one batch[15]. For me, it was easier to download everything in one batch and to work with the dataset locally, as it would avoid causing unnecessary load on the web service.

This full set of court decisions can be downloaded from Rechtspraak.nl and it initially is a .zip file with a size of $\sim$ 5.4GB. When unzipped and unpacked, we get a folder of size 24GB containing **2,890,821**[16] different files. Each file is of the XML (e**X**tensible **M**arkup **L**anguage) type and represents a court decision. This format is both human-readable and computer-readable and it is used to store and transfer data in a structured way [4]. The name of every file is based on the ECLI (section 3.3) corresponding to the particular court decision.

E.g.: the following ECLI translates to the following file name:

"ECLI:NL:HR:1981:AG4158" = ECLI_NL_HR_1981_AG4158.xml

---

[13] https://www.rechtspraak.nl/Uitspraken/paginas/selectiecriteria.aspx

[14] Articles 3, 4 and 5 contain criteria for mandatory publication.

[15] https://data.overheid.nl/en/dataset/uitspraken-rechtspraak-nl

[16] Download: '28-05-2021", the dataset gets updated regularly and so the amount of files in the dataset is susceptible to change.

3.3 ECLI

The European Case Law Identifier (ECLI) is the European standard for the unique numbering of court decisions [23]. In June 2013 the judiciary replaced its precursor, the "Landelijk Jurisprudentienummer (LJN)", by this new standard of labelling court decisions in the Netherlands. With the use of ECLI, references to court decisions should be easier to detect and should therefore, be easier to find [24]. The ECLI consists of the following five parts that are separated by a colon and uniquely identify every court decision.

E.g., 'ECLI:NL:HR:1981:AG4158'

- 'ECLI' : ECLI (self-identifier)

- 'NL' : Country code

- 'HR' : Court identifier

- '1981' : Year of decision

- 'AG4158' : Specific identifier

3.4 File structure

The court decisions in the dataset are of the XML type format, this means that the data in these files is listed in a (tree-like) structured way. In this section I describe the structure and content of these files briefly. For more information about their structure and content, a full technical documentation of these court decisions can be found on Rechtspraak.nl[17].

Generally, court decisions in the dataset follow the structure as seen in figure 4. The structure consists of a root node containing three child nodes; the first child node ('RDF') containing the metadata, the second one ('inhoudsindicatie') containing the abstract and the third one containing either a judgment ('uitspraak') or an advisory opinion[18] ('conclusie').
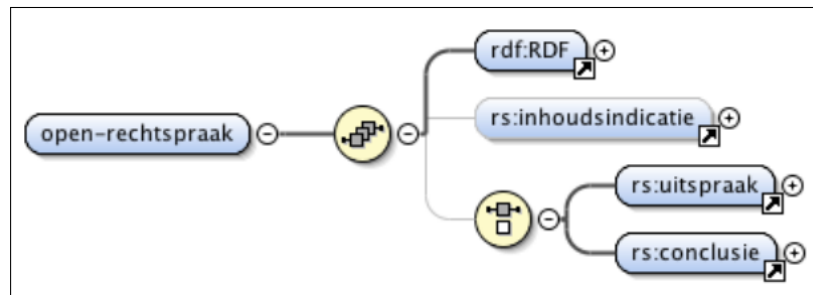

Figure 4: General structure of a court decision file [25]

---

[17]https://www.rechtspraak.nl/SiteCollectionDocuments/Technische-documentatie-Open-Data-van-de-Rechtspraak.pdf

[18]"An advisory opinion is an opinion issued by a court or a commission like an election commission that does not have the effect of adjudicating a specific legal case, but merely advises on the constitutionality or interpretation of a law." – https://en.wikipedia.org/wiki/Advisory_opinion

However, the general structure as seen in figure 4 is not always followed. In all files the metadata section is present, but not all files contain a judgment, advisory opinion or abstract. An overview on the presence of advisory opinions, judgments and abstracts in the files can be seen in table 1.

Section(3)	Abstract	No abstract	Total
Judgment	543,758	178	543,936
Advisory opinion	26,178	2	26,180
'Empty'	0	2,320,705	**2,320,705**
Total	569,936	2,320,885	**2,890,821**

Table 1: Presence of judgments, advisory opinions and abstracts in court decision files

In this table we see that all files with an empty third section miss an abstract section as well. We also see that files containing a judgment or advisory opinion do not necessarily have an abstract. For this project, we are only interested in files that may contain textual references. Therefore, we will only use the files that do not have an empty third section. We are left with: $2,890,821 - 2,320,705 = $ **570,116** files.

N.B. for this project we will call all files that have advisory opinions instead of judgments 'court decisions' as well.
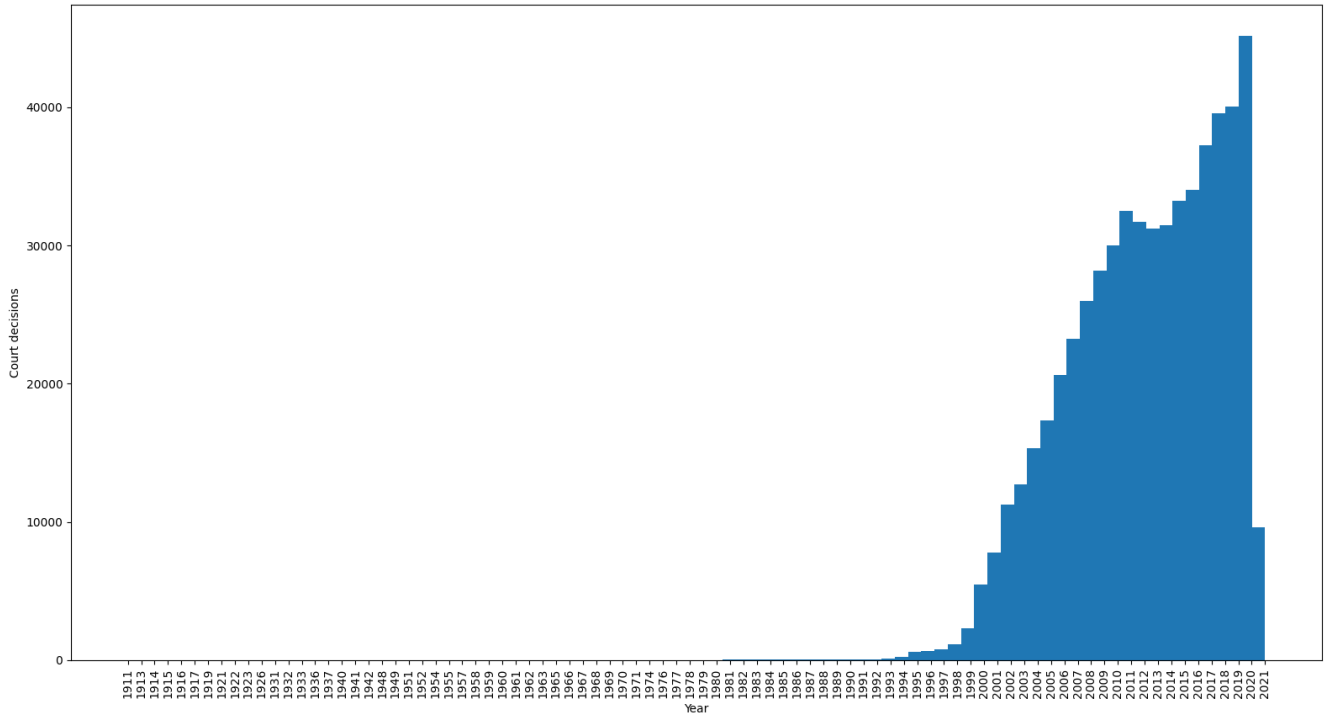
3.5 Dataset characteristics

Figure 5: Court decisions per year [1911-2021]

Figure 5 shows the distribution of court decisions over time. The figure shows that the 570,116 court decisions we are left with have a broad range of 'dates of judgment'. They range from 1911 to 2021 with an overwhelming majority being from 2000 onwards.

It is easy to see that most court decisions in our dataset have been from the last twenty years, this is because Rechtspraak.nl has only been publishing court decisions since its arrival in 1999/2000 (see section 1). Before that time, publication was in the hands of the editorial departments of law reviews. By paying a fee they could opt to receive court decisions from courts in paper and they would have to digitize the decisions themselves [26]. Therefore, it should come as no surprise that courts do not have access to their older court decisions in digital form. However, it is worth noting that the Dutch Supreme Court has been making efforts to digitize its court decisions older than 1999 in the last few years [27].

We also see that the publications per year have been steadily increasing over the years and this trend is likely to continue in the years upcoming. In the NRC article that was mentioned in section 1.1, the chairman of the Council for the Judiciary has stated that the judiciary will aim for a publication rate of 75% in the next 10 years starting from the < 5% it is currently [12]. In my belief, this sounds rather optimistic. However, I do think that this increasing trend will continue, albeit slower than aspired. Interesting to note is that the total amount of cases the judiciary dealt with in 2020 (1,394,860) is lower than in 2019 (1,537,430) [15], yet there are more '2020' cases (45,147) than '2019' cases (40,043) in our dataset. This is probably due to the increased publication rate in 2020 relative to 2019, as was seen in section 1.1 figure 1.

The low amount of publications in 2021 compared to 2020 can simply be explained by the fact that at the moment of writing this thesis, we are still in the middle of 2021. This shows however that the dataset that we are working with is quite up to date.
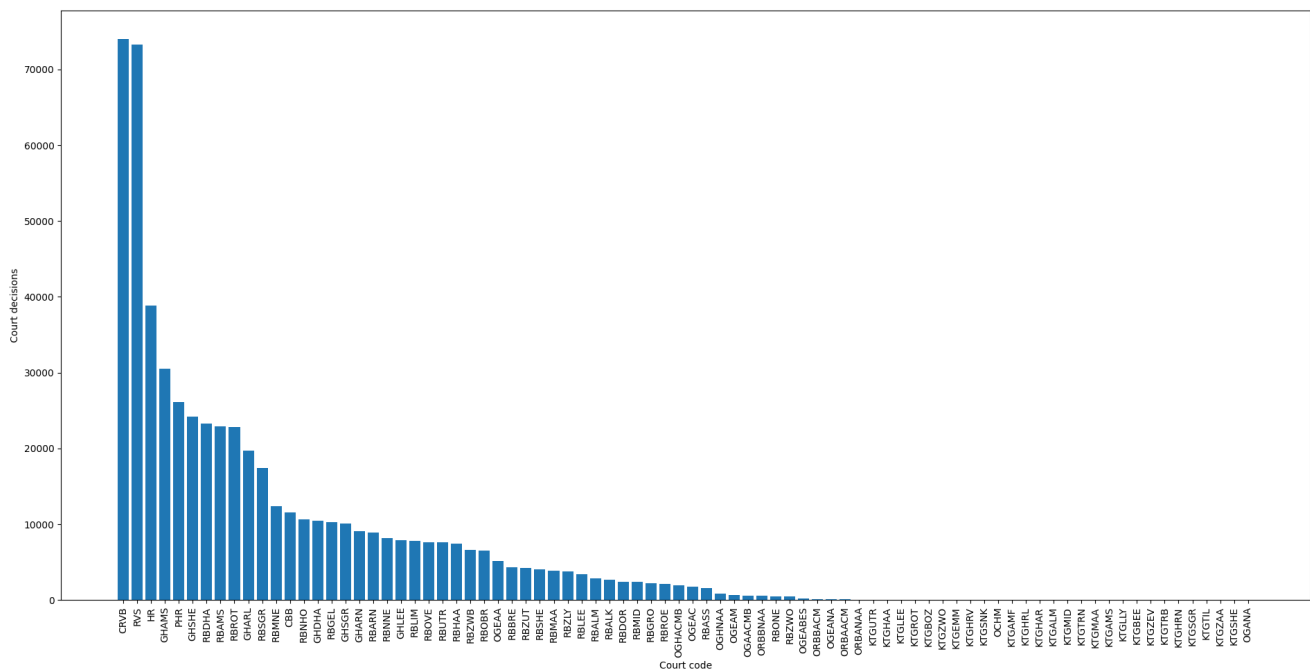

Figure 6: Court decisions per court [1911-2021]

In figure 6 we see the distribution of court decisions per court in our dataset. The representation of every court code can be found on Rechtspraak.nl.[19] In this figure, every judicial district is labelled distinctly. However, if we group all court decisions from the District Courts ('RB') and Courts of Appeal ('GH') together respectively, we can get a clearer view on the distribution as is seen in figure 7.
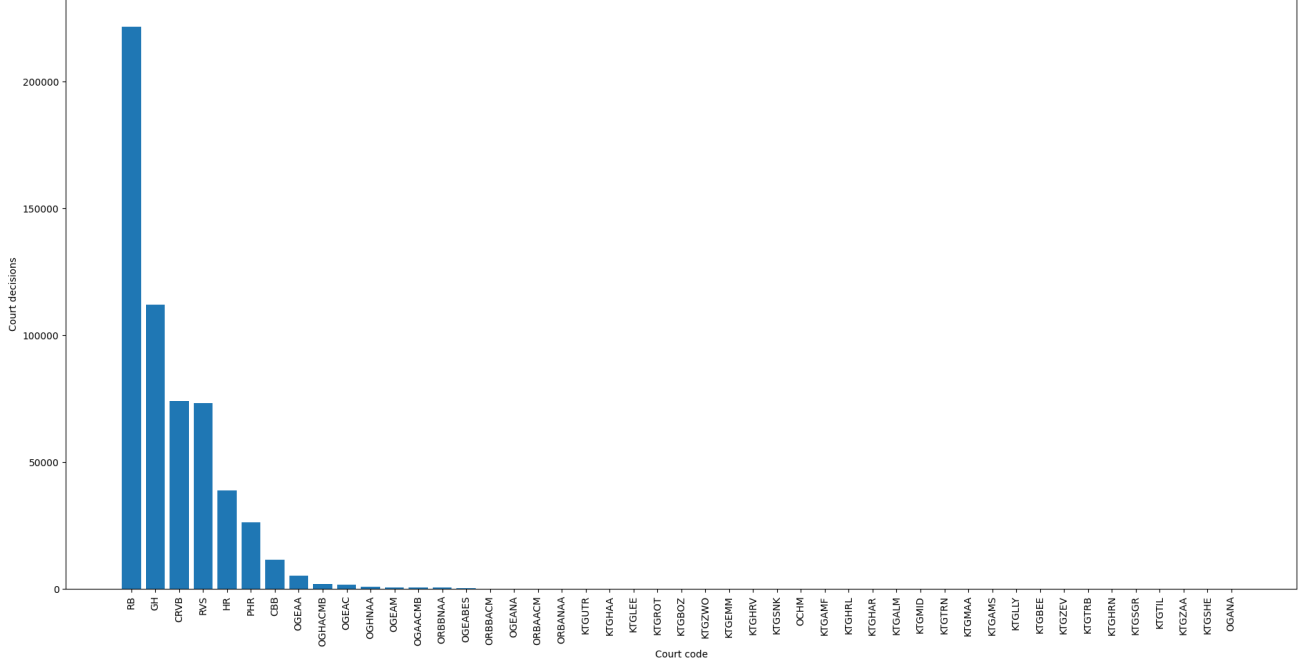

Figure 7: Court decisions per court ('RB', 'GH' grouped) [1911-2021]

We see that the combined amount of decisions from the District Courts and Courts of Appeal make up the majority of our dataset.

Behind them we see decisions from the 'CRvB', 'RvS', 'HR' and the 'CBb' making up a decent fraction of our dataset. Those four courts are the highest courts in the Netherlands for their respective area of law [1].

Figure 7 also shows many different court codes starting with either: 'OC', 'OG', 'OH' or 'OR'. These represent courts from the Netherlands Antilles. The figure also shows many court codes starting with 'KT'; those represent the abolished 'kantongerechten'. Both court decisions from the Netherlands Antilles and the abolished 'kantongerechten' only make up a fraction of the decisions in our dataset.

---

[19]https://www.rechtspraak.nl/Uitspraken/Paginas/Volledige-lijst-Nederlandse-gerechtscodes.aspx

If we dive deeper into the dataset by looking at the distribution of court decisions per court for a single year, we get the following figure:
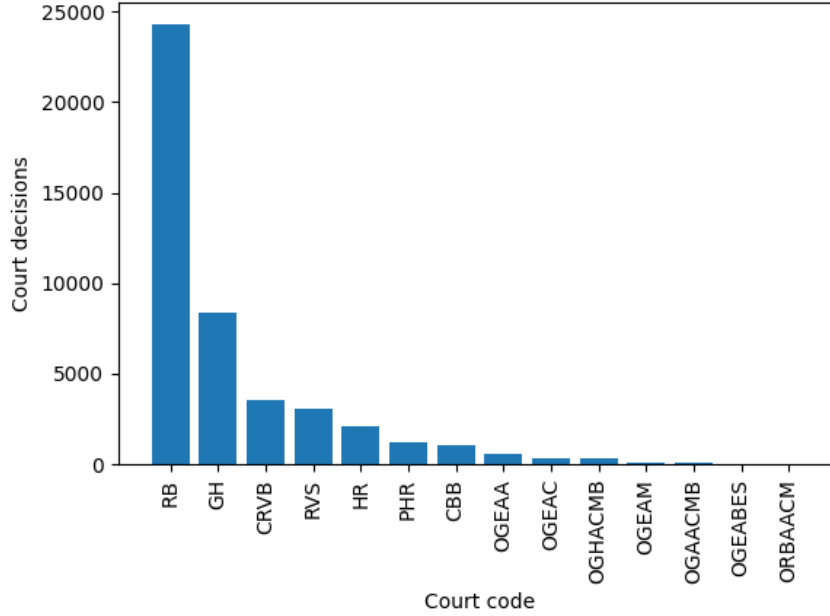

Figure 8: Court decisions per court ('RB', 'GH' grouped) [2020]

In figure 8 we see the distribution of court decisions per court for the year 2020. If we compare the amount of court decisions of the top 7 courts in that figure with the total amount of handled cases per respective court that year, we get the following table:

Courts	Decisions in dataset	Total handled	% of total
District Courts ('RB')	24,321	1,318,500 [15]	≈ **1.84**
Courts of Appeal ('GH')	8,411	41,200 [15]	≈ **20.42**
Central Court of Appeal ('CRvB')	3,511	5,510 [15]	≈ **63.72**
Council of State ('RvS')	3,099	9,393 [28]	≈ **32.99**
Supreme Court ('HR')	2,077	4,505 [29]	≈ **46.10**
The Trade and Industry Appeals Tribunal ('CBB')	1,032	2,340 [15]	≈ **44.10**
Procurator General's Office ('PHR')	1,220	1,480 [29]	≈ **82.43**

Table 2: Court decisions in our dataset vs. total amount of handled cases per court [2020]

The big difference in percentages in table 2 can be explained by how the judiciary selects court decisions eligible for publication on Rechtspraak.nl (see section 3.1). We see that the aim of the selection criteria resolution: "to, in principle, publish all court decisions by the highest courts and to publish the potentially important cases in lower courts" clearly is reflected in these percentages. The highest courts ('CRvB', 'RvS', 'HR', and 'CBB') all have a percentage above 32% whilst the lower courts: 'RB' and 'GH' have percentages of respectively 1.84% and 20.42%.

13

The 'Procurator General's Office' has the highest percentage in table 2, but technically it is not a court. It is part of the same organization as the Supreme Court, however its main job is to provide the Court with expert legal advice (advisory opinions) [29]. It issues the majority[20] of the 'advisory opinions' in our dataset (see table 1).

N.B. the percentages for the highest courts in table 2 would probably be higher if we would compare the court decisions in our dataset against the total amount of handled 'relevant' (see section 1.1) cases.

3.6 Type of references

In the 'schema' documentation found on Rechtspraak.nl, we see that a reference element 'dc-terms:references' can exist in the metadata of our court decisions. The presence of this element indicates that the court decision contains a reference to either: national legislation ('bwb' and 'cvdr'), European legislation ('eu') or jurisprudence[21] ('ecli') [25].

My first step was to try to find references to case law this way. However, if we look at table 3, we see that none of the 227,665 references found, are references to case law; our target. Therefore, another way to extract references to case law was needed (see sections 2.1, 2.2 and 3.7).

Reference type	Count
'ecli'	**0**
'bwb'	227,539
'cvdr'	94
'eu'	32
Total	227,665

Table 3: Distribution of references found (in 86,946 out of the 570,116 files) in our dataset.

Case law references can be made in a lot of ways. To give the reader an indication about the variants of these references, and an indication about what exactly we are looking for; a list of reference variants can be seen in the list below:

- **"ECLI:NL:HR:1998:AA9342"** – This is a reference using 'ECLI', which is the today's standard for citing court decisions [3].

- **"LJN:AA9342"** – Up next we see a 'LJN' reference. LJN was the standard for the unique numbering of court decisions in the Netherlands before the introduction of ECLI in 2013 [24].

- **"NJ 1998, 367"**, **"JM 1998/30"** – We also see references starting with 'NJ' and 'JM', these initials represent two law reviews in which the court decision in question was published. An overview of most Dutch law reviews and their initials can be found here[22].

---

[20]26,161 out of 26,180 by advisory opinions by 'PHR', 3 by 'CBB' and 18 by 'RVS'.

[21]'Case law'

[22]https://documents.library.maastrichtuniversity.nl/open/c2de5acb-006a-4836-860a-eeb1fb721b94

- **"HR, 6 januari 1998, zaaknr 106160 E"** – This next reference is a reference using the following three attributes: the court identifier, the judgment date and the case number[23]. In this instance we see that the decision was issued by the 'HR' = Supreme Court on the 6th of January 1998.

- **"Tweede Pikmeerarrest"**, **"Pikmeer II"** – These last two references are references using aliases, or nicknames. In section 2 of this thesis, these type of aliases were classified as global aliases. These aliases usually take on the name(s) of the parties involved in the case [31].

These references listed, taken from the front page of the LiDO site [32], are all references to the same judicial decision. It must be noted that within these variants, more variations can exist due to e.g. interpunction differences and different word orders.

We see that there are multiple ways to cite one judicial decision, e.g.: ECLI, LJN, law review sources, case attributes and global aliases. However, there are more ways in which court decisions are cited as I have discovered in working on this thesis.

---

[23]All decisions issued by a court get assigned a specific case number [30]

3.7 Linked Data Overheid (LiDO)

A practical example of a case law reference in a court decision file can be seen in figure 9. We see that in the case 'ECLI:NL:RBUTR:2011:BP2283', a reference is made to the case 'ECLI:NL:HR:2008:BD1847', known by many as the "Stoof/Mammoet" case.

```
▼<open-rechtspraak>
  ▼<rdf:RDF xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#" xmlns:ecli="https://e-justice.europa.eu/ecli" xmlns:tr="http://tuchtr
    xmlns:eu="http://publications.europa.eu/celex/" xmlns:dcterms="http://purl.org/dc/terms/" xmlns:bwb="bwb-dl" xmlns:cvdr="http://decent
    xmlns:psi="http://psi.rechtspraak.nl/" xmlns:rdfs="http://www.w3.org/2000/01/rdf-schema#">
    ▶<rdf:Description>
      ...
      </rdf:Description>
    ▶<rdf:Description rdf:about="http://deeplink.rechtspraak.nl/uitspraak?id=ECLI:NL:RBUTR:2011:BP2283">
      ...
      </rdf:Description>
    </rdf:RDF>
  ▼<inhoudsindicatie xmlns="http://www.rechtspraak.nl/schema/rechtspraak-1.0" xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance" xmlns
    id="ECLI:NL:RBUTR:2011:BP2283:INH" lang="nl" xml:space="preserve">
      <para>Geschil tussen werkgever en gewezen werknemer over actuariële herrekening van prepensioen in het licht van de amendementen van
      is in dit geval arbeidsvoorwaarde geworden, die werkgever niet eenzijdig kon wijzigen. Toetsing aan het arrest Stoof/Mammoet.</para>
      </inhoudsindicatie>
    ▶<uitspraak xmlns="http://www.rechtspraak.nl/schema/rechtspraak-1.0" xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance" xmlns:xsd="h
      id="ECLI:NL:RBUTR:2011:BP2283:DOC" lang="nl" xml:space="preserve">
      ...
      </uitspraak>
    </open-rechtspraak>
```

Figure 9: Example of a reference to case law – "ECLI_NL_RBUTR_2011_BP2283.xml"

The XML file itself gives no indication that the text underlined in figure 9 is a reference. Luckily for us, the LiDO website, first mentioned in section 2.1, exists. On this site one can search for judicial decisions and their references. If we type our requested judicial case into the search bar, like in figure 10, we can view the contents of the court decision along with possible references hyperlinked in our web browser. A great feature of this site is that you can search for court decisions using all their reference variants (see section 3.6).

Zoeken   **Brongegevens**   **Help**

Zoeken in LiDO - Linked Data Overheid

Met LiDO - een databank met miljoenen hyperlinks - kunt u snel inzicht krijgen in de verbanden tussen nationale en Europese regelgeving, uitspraken van Nederlandse en Europese rechters, parlementaire documenten en officiële bekendmakingen. LiDO maakt daartoe onder meer gebruik van intelligente software die tal van citatievormen kan herkennen. Als u hieronder bijvoorbeeld zoekt op 'Bolkesteinrichtlijn' dan vindt u documenten waarin verwezen wordt naar deze richtlijn, ook al zijn deze verwijzingen geschreven als 'Richtlijn 2006/123/EG', 'Richtlijn (EU) 2006/123' of 'Dienstenrichtlijn'.

Onder de zoekbox vindt u meer voorbeelden, alsmede een overzicht van de in LiDO opgenomen bronnen. De komende tijd zal het aantal bronnen verder worden uitgebreid en zullen andere functionaliteiten worden toegevoegd. Nu reeds kunt u hier de uitspraken van Rechtspraak.nl raadplegen met toegevoegde hyperlinks. Ook kunt u op eenvoudige wijze zelf betrouwbare hyperlinks maken.

Zoek

ECLI:NL:RBUTR:2011:BP2283

Zoeken   **Wissen**

Figure 10: Search bar on https://linkeddata.overheid.nl

In figure 11 one can see the result of searching for the aforementioned court decision. We see that the reference to the 'Stoof/Mammoet' case is clearly highlighted and clickable. These clickable references should link to the case that is being referenced.

Figure 11: Example of a reference to case law on the LiDO site

By looking at the HTML behind the page when loading the court decision, we see that a get-request is sent involving two components: a link to the court decision on Rechtspraak.nl and a link to the metadata file with extracted references on the LiDO site itself. The request returns the court decision in HTML with the references hyperlinked using the 'href' attribute. The get-request can be seen below:

"https://linkeddata.overheid.nl/front/portal/component/get-document-with-refs-html?id=
http://linkeddata.overheid.nl/terms/jurisprudentie/id/ECLI:NL:RBUTR:2011:BP2283
&document-uri=
https://data.rechtspraak.nl/uitspraken/content?id=ECLI:NL:RBUTR:2011:BP2283"

We already have our court decisions ready in our 'OpenDataUitspraken' dataset, we therefore do not need to retrieve those again. However, we are interested in the metadata file with the extracted references that is found in the first part of the link. An example of what the references in that file look like, can be seen in figure 12.
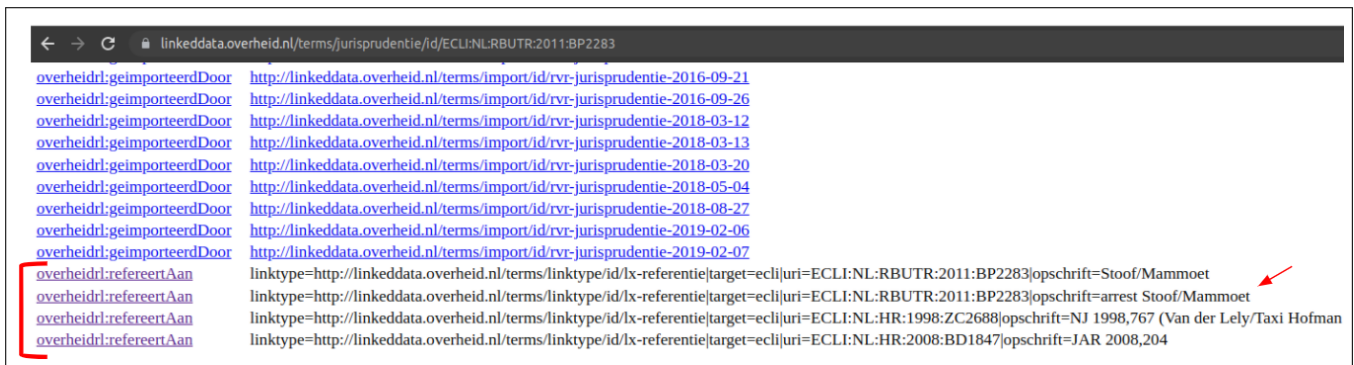
Figure 12: Example of references found on http://linkeddata.overheid.nl/terms/jurisprudentie/id/ECLI:NL:RBUTR:2011:BP2283

In figure 12 above we can see four references, they are marked by the tag 'overheidrl:refereertAan'. Additionally, we see the specific reference that was hyperlinked in figure 11 ("arrest Stoof/Mammoet").

We now have a way to find the references to case law in our court decisions. However, retrieving the metadata files from the LiDO site manually would not be viable considering the amount of court decisions (570,116) we would like to find references in. Subsequently, scraping the LiDO site with a 'webscraper' would probably not be the best idea as well.

We retrieve the metadata by downloading a dump of the LiDO database from:

    https://data.overheid.nl/dataset/linked-data-overheid.

This 38 GB XML file should contain all the metadata files with references to case law that we need for the next step in our research.

4 Methods

In working with the datasets I have mainly used Python [33] (version 3.8.5) however, in the preprocessing parts I have used shell scripts as well. The following Python libraries have been used for general procedures: 'os' [34] for pathing, removing files, renaming files and looping through folders; 'time' [35] for recording run time of code; 'datetime' [36] for storing the download dates of the datasets; 'subprocess' [37] for running scripts from within another script; 'collections.Counter' [38] for extracting the frequencies of elements in list structures.

Libraries that have been used for specific purposes will be described in the sections below accordingly.

4.1 Preprocessing

4.1.1 "OpenDataUitspraken"

As explained in section 3.2, one first has to download the dataset from Rechtspraak.nl. This can be done manually, but in our code it can be done automatically with the use of the 'requests' [39] and 'tqdm'[24] [40] libraries in Python. After unzipping the dataset, one is left with a folder containing 110 subfolders that are named after the year of the court decisions they contain, ranging from 1911 to 2021.[25] Inside these subfolders are up to 12 (one for every month) zipped folders[26] containing all court decisions in question. The unpacking (see section 3.2) consists of unzipping all these nested zipped folders and putting them all into a single folder. I have used shell scripts for the unzipping and unpacking part.

To get an overview of the structure of the court decisions (see section 3.4), I used the Python library BeautifulSoup [41] for traversing the tree-like structures of the court decisions. By looping through all files and checking for the presence of 'uitspraak', 'conclusie' and 'abstract' nodes (see figure 4), I was able to get an overview of the usable files in the dataset (see table 1). Subsequently, removing the 'unusable' files from the dataset was done in a similar way, but with the use of the library ElementTree [42]. Both the BeautifulSoup and the ElementTree libary can be used for parsing and traversing XML files.

4.1.2 LiDO data

After downloading the dump of the entire LiDO database, one has to unzip the downloaded 'lidodata.gz' file. After unzipping it with 'gzip', we are left with a single XML file of size ≈38.2GB. The full database of LiDO can be found in this file.

---

[24]'Visual download bar'

[25]An info.txt file is inside the main folder as well, it however seems outdated and not usable to us.

[26]There is one 'not zipped' folder in the '2016' map that is not supposed to be there, as is confirmed by Rechtspraak.nl (28/05/21 dataset version)

4.2 Reference extraction

The LiDO database gets its data from many sources. It is fed by the xLLx framework that we discussed in section 2.1, and we discovered that it was used for more than just finding references to case law in court decisions. It was also used for e.g.: detecting and solving references to legislation and parliamentary documents. Because of this, the LiDO database also contains lots of data that we have no use for. Luckily, we have already found the type of file that contains the references we are looking for (see section 3.7). Therefore, extracting these files from the LiDO dataset was the next step in our research.

For traversing the single 38GB file, a parser different from BeautifulSoup and ElementTree was needed, as those type of parsers tend to put the entire tree in memory before traversing it. I settled on using the 'lxml.etree.iterparse' [43] library for traversing the LiDO data. This is a parser that parses XML files incrementally and therefore, it is better suited for traversing bigger XML files.

After using the iterparse command to parse the LiDO XML file, we loop through every element in it. An element in an XML file is something that begins with a start-tag: '*<something>*' and ends with an end-tag: '*</something>*'. In between these two tags can be text or other elements. Elements within another element we call nested elements. Within the start-tag, attributes can be present that classify the element in some way; or give extra information about the element e.g., *<something class=else>*. Our court decision files (XML as well) follow the same structure (see figure 9).

In the LiDO data, if we encounter an element that contains the following attribute:[27]

> *{http://www.w3.org/1999/02/22-rdf-syntax-ns#}about*

we know that the element is the start of a node, representing a file.

Most elements do not contain an 'about' attribute, so we skip them. However, if we do encounter an element that has an about attribute, we check if the value of the attribute starts with:

> *http://linkeddata.overheid.nl/terms/jurisprudentie/id/ECLI:NL*

By checking for this specific value, we are confirming that we are getting the metadata files that were seen in figure 12. After encountering such a file, we check if the court decision file belonging to the metadata file is present (in the 'OpenDataUitspraken' dataset). If it is not, we have no use for the metadata file. But if it is, we continue by checking for references in the metadata file.

These references can be found by checking for the following element tag in the metadata file:

> *{http://linkeddata.overheid.nl/terms/}refereertAan*

We then check if there is a text part in this element. Examples of texts found in reference elements were seen in figure 12. If we take the mentioned 'Stoof/Mammoet' reference text from that figure:

> *linktype=http://linkeddata.overheid.nl/terms/linktype/id/lx-referentie|target=ecli|*
> *uri=ECLI:NL:RBUTR:2011:BP2283|opschrift=arrest Stoof/Mammoet*

we can see that it consists of multiple parts; a 'target'; a 'uri'[28] ; and an 'opschrift'. The target refers to the type of reference, e.g.: in a reference to national legislation the target is 'bwb' and in a

---

[27]The text between the curly brackets is a namespace, it can be used for differentiating between XML elements that have the same name.

[28]Coincidentally, the uri in this example refers to the wrong decision.

reference to case law the target is 'ecli' [17]. In a reference to case law, the ECLI of the referenced court decision is stored in the uri part. The final part, 'opschrift', is used for storing the citation itself. In the context of hyperlinks, we call this the anchor text. The anchor text is used for finding the specific reference in the text of court decisions.[29]

We are looking for references to case law, we therefore only want references that have 'ecli' as target. After finding a case law reference, we write the reference to a CSV file using the csv library [44]. Per reference we store three attributes: the ECLI containing the reference (extracted from the 'about' attribute); the target ECLI (uri); and the anchor text (opschrift).

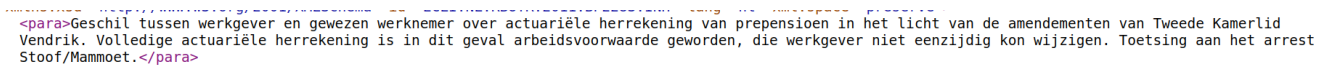We end up with a CSV file containing all our references to case law.

4.3   Error analyses

We have already encountered some errors in the processing step, a list of them will be shown in section 5.2 of this thesis. Additionally, we have also encountered some errors in the extraction part. Those will be listed in section 5.3.

The primary goal of this thesis however, is finding errors in the references extracted. We have seen a list of possible errors in section 2.2 of this thesis. Several types of errors were mentioned and we will describe most of them extensively in the next section. Some of them we will refer to as irregularities. I have used regular expressions in finding these irregularities.

We will give special attention to the false positive type errors in these references, and we will be testing a few methods to find these. To do this we want to take a look at the context of those references. And before we can do that, we need to extract the context.

For each reference in our CSV file, we want to extract the paragraph in which the anchor text was found in the court decision. We extract these paragraphs from the abstract, judgment and advisory opinion sections in the court decision XML files (see section 3.4). These paragraphs are extracted element-wise. In the example below, the anchor text 'arrest Stoof/Mammoet' is found in the 'para' element, thus we want to extract the text[30] in this element.

```
                                                                                                                          preserve
  <para>Geschil tussen werkgever en gewezen werknemer over actuariële herrekening van prepensioen in het licht van de amendementen van Tweede Kamerlid
  Vendrik. Volledige actuariële herrekening is in dit geval arbeidsvoorwaarde geworden, die werkgever niet eenzijdig kon wijzigen. Toetsing aan het arrest
  Stoof/Mammoet.</para>
```

Figure 13: Element containing an anchor text

We will write the reference (ECLI, target ECLI, anchor text) plus the extracted paragraph to a dedicated CSV file. If an anchor text occurs multiple times in one court decision, we add a row to the CSV file for every occurrence of the anchor text in the court decision file.

---

[29]Some references also contain a 'lido-id' part.

[30]Most of the time, a reference is found in the 'element.text' part, but sometimes it is found in the 'element.tail' part.

After writing the paragraph of a reference to a dedicated CSV file, we perform context analysis on the references found. Subsequently, when looping the rows in this CSV file, we use regular expressions to extract all words preceding and following the reference in the paragraph fields. We then find the most common words in those categories. The results of this analysis for a few references can be found in section 5.5. An example of what the CSV files with paragraphs look like, can be seen below:

	A	B	C	D
1	ECLI	Ref_ECLI	Anchor text	Paragraph
2	ECLI:NL:HR:1989:AC3119	ECLI:CE:ECHR:1999:1102JUD003716697	vitale	Binnen die beschouwing past op zijn beurt de strafoverweging "Bovendien heeft de Rechtba
3	ECLI:NL:ORBBNAA:1995:BU4908	ECLI:CE:ECHR:1999:1102JUD003716697	vitale	4.2. Volgens NV X dient op de Nederlandse Antillen een ruimere toepassing te gelden dan ii
4	ECLI:NL:RBROE:1995:1	ECLI:CE:ECHR:1999:1102JUD003716697	vitale	het lichaam en/of het hoofd van voornoemde [slachtoffer 1], waardoor (een) schotverwondin
5	ECLI:NL:RBROE:1995:1	ECLI:CE:ECHR:1999:1102JUD003716697	vitale	hij op 20 januari 1994 te Reuver, tezamen en in vereniging met een ander, opzettelijk [slach
6	ECLI:NL:RBROE:1995:1	ECLI:CE:ECHR:1999:1102JUD003716697	vitale	Er bleek sprake van vier inschot- en één doorschotverwonding. Hierbij waren vitale organen

Figure 14: Example of a CSV file with references and paragraphs stored

Two types of errors that will not be discussed in the next sections are mapping errors and false negatives (see section 2.2). I believe these errors to be harder to detect than false positives, and I have not been able to find methods to detect these.

5 Results

5.1 Extraction statistics

In the last section we have described the process of extracting references from the LiDO data. The LiDO data is supposed to contain all 'linkextracted' court decisions from Rechtspraak.nl [45]. However, we see that we have only found metadata for 486,699 court decisions out of the 570,116 that we have in our 'OpenDataUitspraken' dataset.

This can partly be explained by the fact that the LiDO data we are using, is a database dump that was made in May 2019. The dataset with court decisions that we are using however, is much more recent and it contains court decisions to up until March of this year (2021).

If we look at the $570,116 - 486,699 = 83,417$ court decisions that we did not find metadata for in the LiDO data, and the year of their judgment, we get the following table:

Year	Count
2021	9,621
2020	45,147
2019	27,471
≤ 2018	1,178
Total	83,417

Table 4: Distribution of excluded court decisions based on their year of judgment

The high numbers in 2019, 2020 and 2021 can be explained by the discrepancy in up-to-dateness of both our datasets. But we also see 1,178 court decisions from 2018 and before, that were not found in the LiDO data. A possible explanation for the exclusion of these files is that in the last two years, Rechtspraak.nl has added court decisions from 2018 and before to their dataset.

In the 486,699 court decisions that we had metadata for, only 208,628 court decisions actually had ≥ 1 references to case law in them. In the end, we are left with **638,045** references to case law that were found in 208,628 court decisions.

These 638,045 references are distributed over the courts in the following manner:

Court	Reference count
'PHR'	204,270
'RB'	160,444
'GH'	93,274
'CRVB'	63,624
'RVS'	63,104
'HR'	39,035
'CBB'	10,230
Other	4,064
Total	638,045

Table 5: Distribution of references in court decisions vs. type of court that made the reference

The disproportionately represented PHR is interesting to see as it only made up a relatively small part of the decisions in our dataset (see figure 7). We can only assume that the PHR, or Procurator General's Office, has a habit of frequently citing court decisions in their advisory opinions.

5.2 Formatting errors found in 'OpenDataUitspraken' dataset

The following (possible) errors have been spotted in the 'OpenDataUitspraken' dataset:

- The presence of an outdated info.txt file in the 'OpenDataUitspraken.zip' file.

- The wrongful presence of the '201601' map in the '2016' map in the 'OpenDataUitspraken.zip' file.

- The absence of abstracts in 180 court decision files being inconsistent with the presence of abstracts in the rest of the court decision files (see table 1).

- The absence of 'ecli', 'cvdr' and 'eu' references in comparison to 'bwb' references in the dedicated references section in the metadata part of court decisions (see table 3).

- The court decision file: "ECLI_NL_RBDHA_2016_2379.xml" having an abstract start tag without an abstract closing tag.

- The court decision file: "ECLI:NL:CRVB:1991:ZB4693.xml" having XML metadata added into the judgment part of the file.

These were (possible) errors that I accidentally stumbled upon, many more errors might exist in the 'OpenDataUitspraken' dataset. Finding those however, was not the focus of this thesis.

5.3   Trivial errors found in extraction

In the extraction of references from the LiDO data, we have encountered problems as well. We found 36,557 elements with a reference tag that had no text in them; we ignored those in our research.

The following nine court decisions had a reference in them with a missing anchor text.

- ECLI:NL:RBMAA:2004:AP4284

- ECLI:NL:RVS:2005:AS5505

- ECLI:NL:RBSGR:2009:BK7350

- ECLI:NL:CBB:2013:72

- ECLI:NL:CRVB:2016:2953

- ECLI:NL:GHSHE:2018:1140

- ECLI:NL:RBDHA:2018:5931[31]

- ECLI:NL:GHSHE:2018:5278

- ECLI:NL:RVS:2019:1300

In the metadata of "ECLI:NL:RBMAA:2004:AP4284" on the LiDO site[32], we see that the reference that gave the error is supposed to look like this:

"linktype=http://linkeddata.overheid.nl/terms/linktype/id/lx-
referentie|target=ecli|uri=
ECLI:CE:ECHR:2010:0225JUD003639507 ECLI:CE:ECHR:2008:0918JUD002803404
ECLI:CE:ECHR:2006:1005JUD001255503 ECLI:CE:ECHR:2002:1105JUD004120298
ECLI:CE:ECHR:1997:0317JUD002180293 ECLI:CE:ECHR:2014:0327JUD005496308
ECLI:CE:ECHR:1988:0524JUD001073784
ECLI:CE:ECHR:2013:1008JUD006293012|opschrift=Muller"

In our LiDO data however, it looks like this:

"linktype=http://linkeddata.overheid.nl/terms/linktype/id/lx-
referentie|target=ecli|uri=
ECLI:CE:ECHR:2010:0225JUD003639507 ECLI:CE:ECHR:2008:0918JUD002803404
ECLI:CE:ECHR:2006:1005JUD001255503 ECLI:CE:ECHR:2002:1105JUD004120298
ECLI:CE:ECHR:1997:0317JUD002180293 ECLI:CE:ECHR:2014:0327JUD005496308
ECLI:CE:ECHR:1"

I assume this is because this reference is of such a length, that the opschrift part was cut off in the making of the database dump. This assumption goes for all references with a missing anchor text in the court decisions listed above.

---

[31]The biggest uri we have seen so far can be seen in the LiDO metadata of "ECLI:NL:RBDHA:2018:5931" (169 ECLIs in its uri).

[32]https://linkeddata.overheid.nl/terms/jurisprudentie/id/ECLI:NL:RBMAA:2004:AP4284

5.4 Target reference irregularities

Every reference in our CSV file has a target reference (uri). We immediately see a few irregularities in our references. An overview of them can be seen below:

Target ECLI (uri)	Count
Regular ECLI	568,747
Exclamation mark '!'	25,827
Multiple ECLIs with '!' in front of them	25,096
Multiple ECLIs without '!'	13,010
Empty field ''	5,311
ECLI with '∼' at the end	52
Other	2
Total	638,045

Table 6: Type of uri's found and their formats

We see that only 568,747 references have a single regular ECLI identifier in their target reference field. These are references that have been correctly detected and resolved.

If the target reference field has an exclamation mark in it, it means a reference has been correctly detected but the document is not found in the LiDO database, and so the reference can not be resolved. The LiDO site shows this explanation when searching for the anchor text of such a reference on the site, it also shows that a spelling error in the reference may be the cause. These references fall under the "incomplete or erroneous reference" type errors as mentioned in section 2.2. A top 100 of most common anchor texts for these references can be found in appendix A.

We also see 25,096 references that contain multiple ECLIs with exclamation marks in their uri. Most anchor texts in these references look like case attributes, LJN's or law review sources (see section 3.6). A typical example of these references can be seen below:

linktype=http://linkeddata.overheid.nl/terms/linktype/id/lx-referentie|target=ecli|
uri=!ECLI:NL:HR:1994:BV4750!ECLI:NL:HR:1994:ZC5712|
opschrift=Hoge Raad van 13 juli 1994, nr. 28 997

In the example above, the anchor text refers to a judicial decision using case attributes. If we take a deeper look at both ECLIs in the uri, we see that they are both issued by the Supreme Court on the 13th of July 1994. They even have a very similar case number, '28 997' vs. '28997'. It is not surprising that it is unclear to which of these two decisions is referred. Out of the 25,096 references in this category, 19,649 references have multiple ECLIs in their uri that are issued by the same court in the same year.

Additionally, we see 13,010 references with multiple ECLIs in their uri, but without the exclamation marks. E.g.:

linktype=http://linkeddata.overheid.nl/terms/linktype/id/lx-referentie|target=ecli|
uri=ECLI:CE:ECHR:2015:1215JUD002902411 ECLI:CE:ECHR:2000:1017JUD004505998|
opschrift=bono

In this example, the anchor text 'bono' leads to the existence of two ECLIs in the uri. This is not surprising as the first ECLI translates to the 'BONO v. FRANCE' case and the second ECLI translates to the 'BONO v. ITALY' case. This is an instance of two judicial decisions having the same alias and the extractor not knowing which one is referred to. Most of the references in this category seem to have anchor texts that are aliases.

I believe both types of 'multiple ECLI' references can be classified as the 'ambiguity' type errors mentioned in section 2.2. However, both categories are ambiguities for different reasons. A top 100 of most common anchor texts for these ambiguous references can be found in appendix B.

Up next are the 5,311 references with no uri at all. Most of these references seem to have case aliases as their anchor text. Additionally, as opposed to the ambiguous and incomplete/erroneous references, the references with empty uri's seem to be hyperlinked properly on the LiDO site. An example can be seen below:

```
linktype=http://linkeddata.overheid.nl/terms/linktype/id/lx-referentie|target=ecli|
uri=|opschrift=Carson
```


We also have 52 references that have a '∼' at the end of their uri. E.g.:

```
linktype=http://linkeddata.overheid.nl/terms/linktype/id/lx-referentie|target=ecli|
uri=ECLI:CE:ECHR:2003:1113JUD002314593∼
|lido-id=|opschrift=ELCI
```

I have no explanation for the presence of the '∼' character at the end of the uri in these references. However, I do observe that all ECLIs in the uri's of these references are judicial decisions by the ECHR, the European Court of Human Rights.

The two 'Other' references have obvious errors in their uri:

```
linktype=http://linkeddata.overheid.nl/terms/linktype/id/lx-referentie|target=ecli|
uri=<em class="attr"></em>ECLI:NL:HR:2005:AT9061|
opschrift=Nap/Rabobank
```

Somehow a part of XML code has slipped into the uri in the example above.

```
linktype=http://linkeddata.overheid.nl/terms/linktype/id/lx-referentie|target=ecli|
uri=CLI:NL:HR:2001:AB0379
|opschrift=Bührmann/Celtona
```

In this example, the uri starts with 'CLI' instead of 'ECLI'.

More irregularities might exist within the 568,747 'regular ECLI' references set, I have not checked all of them. We assume them to be correct.

5.5 Anchor text irregularities

In the anchor texts of our 638,045 references we do not immediately see any irregularities. I see 364 empty anchor texts and these are assumed to be erroneous.

What I did next was to look at the top 100 most common anchor texts for all references. The list can be found in appendix C. The top 10 can be seen in the table below:

Anchor text	Count
'conclusie'	8,611
'Cichowski'	2,907
'vitale'	1,813
'Burg'	1,282
'boden'	1,148
'Salduz'	952
'dagdelen'	915
'Bosnak'	835
'Jurgens'	814
'Muller'	783
Other	617,070
Total	638,045

Table 7: Top 10 most common anchor texts in court decisions

I assume that even though, 'conclusie' is detected as the most common anchor text for references in the LiDO data, the exception list mentioned in section 2.1 prevents the anchor text from being hyperlinked on the LiDO site itself. Consequently, I have not been able to find the word 'conclusie' highlighted on the LiDO site in any of the court decisions I checked.

We also see the Dutch words: 'vitale', 'boden' and 'dagdelen' in this top 10. I assume these anchor texts are aliases of judicial decisions however, they do raise suspicion of being false positives as they occur quite frequently in our references while being common Dutch words. These suspicions where confirmed when I found the following instances of those words being incorrectly detected as references:

. De verdachte heeft aangever in de buurt van <u>vitale</u> organen gestoken, wat zeer ernstig letsel op kan leveren.

. de verschillende <u>dagdelen</u> waarop de horeca wordt toegestaan,

gesprekken met zijn oom en bezoek(en) aan de huiskerk. Verweerder had nader moeten motiveren waarom hij desondanks eisers verklaringen op dit punt niet volgt. Met de tegenwerping van verweerder dat het voorstelbaar is dat de gesprekken met zijn oom eiser troost <u>boden</u>, maar dat dit nog niet wil zeggen dat dit voor hem aanleiding vormde zich

In the examples above, one can see the anchor texts being used as words in regular Dutch sentences. There is no indication whatsoever that a reference is made to a judicial decision in these sentences. Therefore, I believe that these references are incorrectly recognized as references by the xLLx (see section 2.1). I suspect that either these anchor texts are incorrectly recognized as global aliases by the 'NER' component, or they have been mistakenly left in by the 'false-positive removal' component.

If we look at the most common words following a 'vitale' reference in our court decisions, we get the following table:

Word following 'vitale'	Count
'organen'	805
'delen'	515
'lichaamsdelen'	232
'kenmerken'	124
'structuren'	95
'functies'	72
'rol'	62
'onderdelen'	55
'depressie'	50
'en'	40

Table 8: Top 10 most common words following a 'vitale' reference

The table above strongly indicates that the 'vitale' anchor text is very prone to being incorrectly recognized as a false positive. This is because the words in the table make up common adjective-noun combinations with the word 'vitale' in Dutch, e.g.: 'vitale organen' = vital organs.

If we look at the areas of law of the 1,813 court decisions in which the 'vitale' case was referenced, we unsurprisingly see that the reference is most commonly seen in criminal law cases:
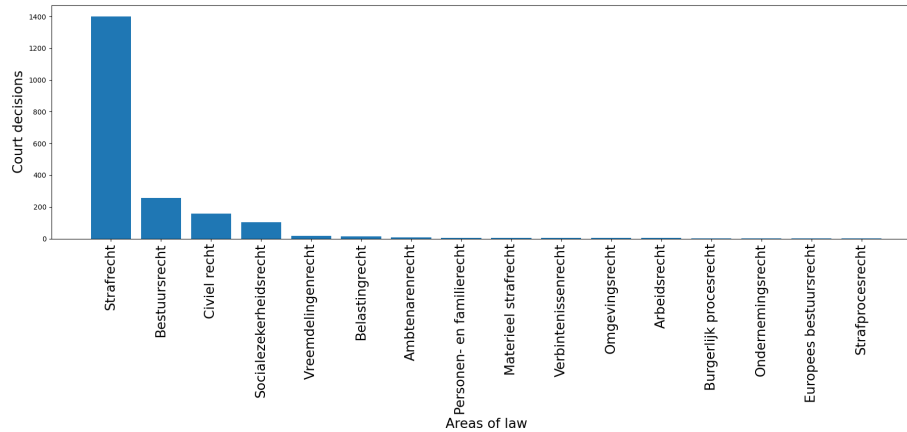

Figure 15: Areas of law distribution of cases in which 'vitale' was referenced

If we look at the most common words following the anchor text 'boden', we get words like: 'voor', 'de', 'om', 'en', 'op', etc. These words can, in principle, be used behind a lot of words. However, if we look at the most common words preceding the word 'boden', we get the following table:

Word preceding 'boden'	Count
'grondslag'	80
'ruimte'	43
'bescherming'	36
'aanknopingspunten'	35
'grond'	30
'zekerheid'	29
'mogelijkheid'	28
'verhaal'	28
'en'	25
'dekking'	21

Table 9: Top 10 most common words preceding a 'boden' reference

We see a lot of words that can only precede a limited number of verbs in Dutch. The Dutch verb 'boden' being one of them.

Depending on the type of the word, taking the word preceding the anchor text might be a better indication of the reference being a false positive than the word following the anchor text.

Another interesting observation is that the anchor texts 'Cichowski', 'Bosnak', 'Jurgens' and 'Burg' sometimes are falsely detected as references. In the examples below it is apparent that those anchor texts refer to a person, and not the judicial decision. This problem occurs when a person involved in the case has the same name as the alias of a court decision.

> Dit arrest is gewezen door de vice-president D.G. van Vliet als voorzitter, en de raadsheren E.N. Punt en P.M.F. van Loon, in tegenwoordigheid van de waarnemend griffier E. Cichowski, en in het openbaar uitgesproken op 7 mei 2010.

> wonende te [plaatsnaam] ,
> appellanten,
> advocaat: mr. J.M. Bosnak,
>
> tegen:

> Aldus vastgesteld door mr. G.T.J.M. Jurgens, lid van de enkelvoudige kamer, in tegenwoordigheid van mr. M.C. van Engelen, griffier.

in tegenwoordigheid van mr. A.E. van der Burg, griffier,

When we check all paragraphs where a reference to 'Cichowski' is made in our court decisions, we see that in every paragraph either the word: 'griffier', 'rechter' or 'voorzitter' occurs. These are all job titles, the Cichowski reference never occurs in other contexts. Therefore, it is doubtful if any true reference to the Cichowski case exists in our data.

In the top 10 of most common anchor texts in table 7, we also see the 'Salduz' case. I believe this to be the only anchor text in the top 10 that does not have many false positives. If we look at the most common words preceding and following 'Salduz' references, we get the following table:

Word preceding 'Salduz'	Count	Word following 'Salduz'	Count
'de'	636	'tegen'	247
'het'	232	'en'	105
'zaak'	132	'v'	90
'inzake'	128	'jurisprudentie'	80
'gevoerde'	81	'verweer'	67
'zogenoemde'	70	'versus'	35
'zogenaamde'	63	'vs'	33
'zaken'	58	'arrest'	25
'van'	53	'voorschriften'	12
'een'	40	'27'	11

Table 10: Top 10 most common words preceding and following a 'Salduz' reference

We see that the words around 'Salduz' references are different from the words in the previous tables, e.g.: 'arrest', 'zaak', 'zaken' and 'jurisprudentie'. Those words translate to respectively: judgment, case, cases and jurisprudence. These are all words one could expect around a reference to case law. Therefore, I expect the Salduz references to be mostly true positives. It might be interesting to compare the context of true positive references with the context of 'possible' references for the detection of false positives references.

We see that the top 10 most common anchor texts in our references are prone to being false positives. Therefore, being a common anchor text might be in itself an indicator for being false positive. However, the only way to falsify these references with 100% certainty is through checking them manually.

5.6   URI statistics

In section 5.4 we listed the most common anchor texts in our references. However, if we take a look at the top 100 (see appendix C), we see that multiple anchor texts can refer to the same judicial case, e.g.: "Haviltex-maatstaf", "Haviltex", "Haviltex-criterium", "Haviltexmaatstaf" all refer to one judgment.

If we take a look at the top 10 most common uri's in the set of 568,747 references with regular uri's (see section 5.4), we get a better indication of the most commonly referenced court decisions:

ECLI in uri	Case alias	Count
'ECLI:NL:HR:1981:AG4158'	Haviltex	3,795
'ECLI:CE:ECHR:1999:1102JUD003716697'	CASE OF VITALE AND OTHERS V. ITALY	1,961
'ECLI:CE:ECHR:2003:0128DEC003476302'	BURG and OTHERS v. FRANCE	1,412
'ECLI:NL:HR:2008:BD2578'	Redelijke termijn II	1,235
'ECLI:CE:ECHR:1987:1027JUD001093084'	CASE OF BODÉN v. SWEDEN	1,147
'ECLI:NL:HR:2005:AO9006'	-	988
'ECLI:CE:ECHR:2008:1125JUD000176703'	CASE OF DAĞDELEN ... v. TURKEY	937
'ECLI:NL:XX:2008:BH0402'	Salduz	916
'ECLI:NL:CRVB:2009:BH1009'	-	839
'ECLI:CE:ECHR:2007:0828DEC001387605'	BOSNAK v. SLOVAKIA	835

Table 11: Top 10 most common uri's in court decisions

We see the well known 'Haviltex' case at the top. We also see many ECLIs corresponding to the anchor texts seen in table 7. Even with my limited knowledge of the law, I find the abundance of ECHR cases in this list suspicious. Especially, because parts of their aliases are common Dutch words if we ignore punctuation marks and uppercase letters. And additionally, because in section 2.3, we have seen that ECHR cases were not commonly referenced court decisions. However, those statistics were from a decade ago and a lot might have changed. A top 100 of the most common uri's in court decisions can be found in appendix D.

These ECHR decisions themselves can be found in a database dedicated to ECHR case law named 'HUDOC'. In this database every decision is ranked from 1 to 3 in importance level. A ranking of 1 indicates that the case is of high importance; a ranking of 2 indicates medium importance; and a ranking of 3 indicates low importance [46]. If we take a closer look at the referenced ECHR decisions above and their importance level, we get the following table:

Case alias	Importance level
CASE OF VITALE AND OTHERS V. ITALY	3
BURG and OTHERS v. FRANCE	(Key case)
CASE OF BODÉN v. SWEDENN	2
CASE OF DAĞDELEN AND OTHERS v. TURKEY	3
BOSNAK v. SLOVAKIA	3

Table 12: ECHR court decisions and their importance level

We see that three out of five decisions are marked as decisions of low importance. In the HUDOC FAQ [47], we see that this rating translates to: "Judgments, decisions and advisory opinions of little legal interest, namely judgments and decisions that simply apply existing case-law, friendly settlements and strike outs (unless raising a particular point of interest)."

It is therefore interesting to see that in our Dutch court decisions, these ECHR cases of seemingly low importance are cited as often as they are.

5.7 Future references

Another observation that we have made is that, in some cases, the year of judgment of a referenced case is in the future, e.g.:

ECLI	Ref_ECLI	Anchor text
ECLI:NL:HR:1977:AB7174	ECLI:CE:ECHR:2014:0918JUD002101010	Brunet

Initially, one might expect all 'future references' to be false positives since it seems impossible to refer to a decision that has not happened yet. However, exceptions do occur as can be seen below:

ECLI	Ref_ECLI	Anchor text
ECLI:NL:RBGRO:2001:AB3331	ECLI:NL:CRVB:2003:AN8380	Centrale Raad van Beroep bevestigd op 4 november 2003, 01/1907

In the case above, it seems like the reference has been edited in on a later date. Nevertheless, I believe the majority of these future references to be false positives. If we look at the amount of future references in the set of 568,747 references with regular ECLIs as uri, we get the following distribution over referenced courts.

Court	Future references	Total references	% of total references
ECHR	4,288	43,322	$\approx$ **9.90**
GH	315	32,843	≈ 0.96
CRVB	148	54,391	≈ 0.27
HR	146	246,562	≈ 0.06
RVS	130	71,713	≈ 0.18
RB	128	57,352	≈ 0.22
C (EU Court of Justice)	99	36,061	≈ 0.27
Other	18	26,503	≈ 0.07
Total	5,272	568,747	≈ 0.93

Table 13: Future references (≥ 1 year difference) and type of court that is referenced

We see that ECHR references are disproportionately represented in the table and that **9.90%** of all ECHR references in the dataset seem to be future references. This reaffirms our suspicion that ECHR references are prone to being false positive.

5.8 Missing references

Something to note is that not all anchor texts are found in their respective court decisions. An example of why this occurs, can be seen below:

> linktype=http://linkeddata.overheid.nl/terms/linktype/id/lx-referentie|target=ecli|
> uri=ECLI:NL:RBAMS:2005:AU2950
> |opschrift=LJN AU2950

. Het vonnis is gepubliceerd onder LJN AU 2950.

We see that the reference has not been highlighted in the text on the LiDO site because in the reference, the anchor text is "LJN AU2950" and in the court decision itself it is "LJN AU 2950". There is a space between 'AU' and '2950' and therefore, the anchor text is not found in the court decision.

Out of the 638,045 references 610,175 (95.63%) were found in the court decisions.

6   Discussion

We have seen multiple types of errors in our data. Many of which have been mentioned before in section 2.2 of this thesis. We extensively analysed the uri's and anchor texts of our references, and we have looked at the context of these references. By doing this we have found ambiguities, incomplete/erroneous references and false positives.

In analysing the most common anchor texts in our references, we have seen that some of them are common Dutch words. We have analysed the context of these Dutch words and we have showed that it is likely that they do not refer to judicial decisions; they are false positives. The believed cause for the presence of these false positives is an error in the xLLx extraction process (see section 2.1). The anchor texts are seen as global aliases of a case and so they are incorrectly labelled as named entities, and by extent recognized as references. The 'false positive removal' component of the xLLx then fails in filtering these false positives later on in the pipeline.

Therefore, we can choose to improve the named entity recognition component of the xLLx, so that we can reduce the amount of false positives that are marked as named entities. Or, we can improve the 'false positive removal' component so that it is better at filtering the false positives in the data.

My first suggestion for this problem would be to have the label of global alias assigned more critically, especially for possible references to ECHR cases. These cases usually have an alias that is of the following structure: "Person x vs. Country y". I believe that the majority of false positives in our data exists because global alias labels were given to texts that solely matched on the "Person x" part instead of the whole structure of the alias. For well known cases (e.g., Salduz) this should be fine however, I believe we would get better results if less known cases would only match on the entire alias. One way to decide upon the strictness of matching could be to take the importance level rating of the ECHR case into account. (Low rating = less known = stricter matching, if we assume that judges are more familiar with important cases than unimportant cases).

Another option could be to use context analysis. If the xLLx were to look at the words around a possible reference, it might figure out early if a reference is a true positive or not. Words like: "arrest", "jurisprudentie" and "zaken" around a reference might indicate that a reference to case law truly is made. Van Opijnen has already implemented a method for this as mentioned in section 2.2. Consequently, I believe he should add many more aliases to the list of aliases that have to appear in combination with other words. Including: all aliases of ECHR cases that are also common Dutch words (see section 5.5); all aliases of ECHR cases that have a low importance rating (see section 5.6); or maybe just all ECHR cases in general.

Another problem we have encountered is that some common anchor texts are aliases of court decisions, while they are also names of people involved in the court decision. To avoid this mix-up, we could have the 'pattern recognition' component recognize personal names in court decisions and exclude them from being classified as references. Any previously detected reference in the personal names would be dismissed.

All in all, its all about balancing the false positives and false negatives in the data. If we were to assess possible references more critically, less texts would be seen as references and we could decrease the false positive cases in our data. Yet, it could also lead to us missing some of the actual references and this would subsequently lead to more false negatives in the data. In my belief, it is

acceptable to have more false positives than false negatives in the data, as I think it is easier to find and correct false positives than it is to find and correct false negatives.

A more sophisticated method to analyse the context of references has not been discussed yet. 'Word embeddings' are vector representations of words using the context around them (neighbour words) [48]. It is possible to measure the similarity between vectors and this similarity can be represented numerically. If we could create a vector of true positive references, we could measure the similarity between this vector and the vector representation of a 'possible' reference. A high similarity could indicate that a possible reference is a true reference and a low similarity could, per contra, indicate that it is not a reference at all. To model this true positive vector, we would need labelled data (supervised learning) which we do not have. However, we could train the model using data with 'assumed' labels. For example, references made using ECLI can relatively safely be assumed to be correct. Training a true positive model using these ECLI references would be a type of weakly supervised learning. The effectiveness of using these true positive vectors for finding false positives is unknown as this method has not been extensively explored in this thesis. It might be worth researching in a future project.

7 Conclusions and further research

We have now reached the end of this thesis. Our goal was to analyse extracted case law references in Dutch court decisions. I believe that we have got a good look at all the parts that made up our goal. We have extensively analysed the Dutch court decisions that are available to us. We have explained: why they are available to us; which ones are available to us; and we have explained all their characteristics. Additionally, we have described: the role of case law in our legal system; the ways of referencing case law; and the ways of finding these references. Consequently, we have taken a look at the software behind the extraction of the references and we have seen what can go wrong in this process. Furthermore, we have also discussed the options for optimizing the software to reduce the amount of false positive references it extracts.

Alias references to ECHR court decisions seem to be the cause of most false positives in our data. We have suggested several methods to combat these false positives from within the extraction process, hopefully it is feasible to put these methods into practice. Nevertheless, I expect that in future court decisions, the problem solves itself as ECLI has become the standard of citing court decisions.

One might also consider the use of word embeddings in finding false positive references (see section 6). The use of word embeddings for this cause would require us to have extracted references that are labelled. One way to get labelled data could be to have the general public label our extracted references. In the example program below, a random anchor text and the paragraph it is found in are shown. The user of the program then has to judge if the anchor text in the paragraph is a true reference or not. Facilitating such a program on a public website (e.g., LiDO) could be a way to generate labels for our references. However, for optimal accuracy, one might have to consider limiting its target audience (e.g. law students or legal professionals).

```
Total:  0
False positives:  0
True positives:  0

Anchor text:  vitale

Sentence:  Binnen die beschouwing past op zijn beurt de strafoverweging "Bovendien heeft de Rechtbank het de verdachte bijzonder kwalijk genomen, dat hij na een geschil om
 een kleine hoeveelheid cocaine, niet heeft geaarzeld om het slachtoffer herhaaldelijk en met kracht steekwonden toe te brengen in vitale delen van het lichaam".

Is this a true positive, type: `y'

Answer: ▮
```

Figure 16: Example program for data annotation

Additionally, we have seen that the judiciary aims to publish 75% of its handled cases (see section 3.5). This comes down to the publication of ≈1.1 million cases per year if we take the amount of cases that were handled in 2020. That is almost twice the amount of decisions that we were able to use for this thesis (570,116). This massive increase of available data in the next decade would probably lead to more interesting applications of extracted references to case law.

"Checking and possibly improving the links found –
e.g. to process the 'resolve manual errors' –
is a task for end-users" – **Opijnen, 2015** [17]

37

References

[1] R. voor de rechtspraak, Rechtspraak in nederland, Brochure (2019).
URL https://www.rechtspraak.nl/SiteCollectionDocuments/
R002-Rechtspraak-in-Nederland.pdf

[2] Uitspraken (2021).
URL https://www.rechtspraak.nl/Uitspraken

[3] M. van Opijnen, Alle uitspraken online? hoe dan?, Ars Aequi 01/02/21 (2021) 128,131.

[4] M. van Opijnen, Op en in het web: Hoe de toegankelijkheid van rechterlijke uitspraken kan
worden verbeterd, UvA-DARE, 2014.

[5] https://libguides.ru.nl/rechten/nederlandse_jurisprudentie_vinden (2021).
URL https://libguides.ru.nl/rechten/nederlandse_jurisprudentie_vinden

[6] T. Barkhuysen, Een wettelijke regeling voor publicatie van rechterlijke uitspraken op
rechtspraak.nl, Nederlands Juristenblad 24 (2017) 1649.

[7] J. M. de Meij, A. H. Ekker, J. F. Haeck, A. W. Hins, L. W. M. Wopereis, Toegang tot
rechtelijke uitspraken: rapport van de vmc-studiecommissie openbaarheid van rechtspraak.,
Mediaforum 18(4, bijlage) (2006) 121–142, page 3 in article itself, page 34 in article itself.

[8] What is the difference between common law and civil law? (2021).
URL https://onlinelaw.wustl.edu/blog/common-law-vs-civil-law/

[9] A. E. P.B. Cliteur, Inleiding recht, Wolters Kluwer, 2013.

[10] Hoe-komt-een-rechter-tot-uitspraak (2021).
URL
https://www.rechtspraak.nl/Organisatie-en-contact/Rechtspraak-in-Nederland/
Rechters/Paginas/Hoe-komt-een-rechter-tot-uitspraak.aspx

[11] M. van Opijnen, A model for automated rating of case law, ICAIL 2013 (2013) 5.2.3.

[12] S. Derix, Baas rechtspraak: 'misschien hebben we te weinig gesproken over de uitvoerbaarheid
van bepaalde wetgeving', NRC (2021).
URL https://www.nrc.nl/nieuws/2021/05/30/
baas-rechtspraak-de-rechtsstaat-staat-weer-op-de-agenda-a4045430

[13] Verstek, verstekvonnis en verzet (2021).
URL https:
//www.amsadvocaten.nl/de-rechtsgebieden/procesrecht/verstekvonnis-en-verzet/

[14] De rechtspraak, kengetallen 2019, Tech. rep., Raad voor de rechtspraak (2019).

[15] De rechtspraak, jaarverslag 2020, Tech. rep., Raad voor de rechtspraak (2020).

[16] M. D. (Rechtsbescherming), Vragen gesteld door de leden der kamer, met de daarop door de regering gegeven antwoorden: 2255 (2019-2020).

[17] M. Opijnen, N. Verwer, J. Meijer, Beyond the experiment: the extendable legal link extractor, 2015, pp. 1,2,3,4,5,6,8.

[18] M. Opijnen, Canonicalizing complex case law citations., 2010, pp. 97–106. doi:10.3233/978-1-60750-681-2-97.

[19] M. van Opijnen, Search engines don't understand lawyers, but smart technologies can improve access to legal information, European Central Bank ESCB Legal Conference 2018 (2018).

[20] R. Winkels, J. Ruyter, Kroese, Determining authority of dutch case law, Frontiers in Artificial Intelligence and Applications 235 (12 2011). doi:10.3233/978-1-60750-981-3-103.

[21] E. Maat, R. Winkels, T. Engers, Automated detection of reference structures in law, 2006, pp. 41–50.

[22] Uitspraken rechtspraak.nl (2021).
URL https://data.overheid.nl/en/dataset/uitspraken-rechtspraak-nl

[23] Ecli (en ljn) (2021).
URL https://www.rechtspraak.nl/Uitspraken/paginas/ecli.aspx

[24] M. van Opijnen, Rechtspraakdata: open, linked en big, Rechtstreeks 2014 nr 2 (2014) 16.

[25] Schema documentation for open-rechtspraak.xsd (2012).
URL https://www.rechtspraak.nl/Uitspraken/Paginas/Open-Data.aspx

[26] T. Scheltema, Oude arresten niet vrij toegankelijk (2018).
URL https://www.advocatenblad.nl/2018/06/04/oude-arresten-niet-vrij-toegankelijk/

[27] Eerste 150 'oude' arresten gepubliceerd (2021).
URL https://www.rechtspraak.nl/Organisatie-en-contact/Organisatie/Hoge-Raad-der-Nederlanden/Nieuws/Paginas/Eerste-150-oude-arresten-gepubliceerd.aspx

[28] Bestuursrechtspraak in cijfers, jaarverslag 2020, Tech. rep., Raad van State (2020).
URL https://jaarverslag.raadvanstate.nl/2020/wp-content/uploads/sites/9/2021/03/Jaarverslag-2020-Bestuursrechtspraak-in-cijfers.pdf

[29] Hoge raad der nederlanden jaarverslag 2020, Tech. rep., Hoge Raad der Nederlanden (2020).
URL https://2020.jaarverslaghogeraad.nl/2020-highlighted/preface/

[30] Zoeken op ecli of ljn, zaak-/parketnummer of vindplaats. (2021).
URL https://www.rechtspraak.nl/Uitspraken/Paginas/Hulp-bij-zoeken.aspx

[31] Leidraad voor juridische auteurs : voetnoten, bronvermeldingen, literatuurlijsten en afkortingen in Nederlandstalige publicaties, negende druk. Edition, Wolters Kluwer, Deventer, 2019.

[32] Lido frontpage (2021).
URL https://linkeddata.overheid.nl/front/portal/

[33] Python (2021).
URL https://www.python.org/

[34] Python: os (2021).
URL https://docs.python.org/3/library/os.html

[35] Python: time (2021).
URL https://docs.python.org/3/library/time.html

[36] Python: datetime (2021).
URL https://docs.python.org/3/library/datetime.html

[37] Python: subprocess (2021).
URL https://docs.python.org/3/library/subprocess.html

[38] Python: Counter (2021).
URL https://docs.python.org/3/library/collections.html

[39] Python: requests (2021).
URL https://docs.python-requests.org/en/master/

[40] Python: tqdm (2021).
URL https://tqdm.github.io/

[41] Python: BeautifulSoup (2021).
URL https://www.crummy.com/software/BeautifulSoup/bs4/doc/

[42] Python: ElementTree (2021).
URL https://docs.python.org/3/library/xml.etree.elementtree.html

[43] Python: lxml.etree.iterparse (2021).
URL https://lxml.de/api/lxml.etree.iterparse-class.html

[44] Python: csv (2021).
URL https://docs.python.org/3/library/csv.html

[45] M. van Opijnen, Over de noodzaak tot betere standaardisatie van juridisch bronmateriaal, Computerrecht 2018/51 (2018).

[46] Hudoc user manual, Tech. rep., ECHR (2021).
URL https://www.echr.coe.int/Documents/HUDOC_Manual_ENG.PDF

[47] Hudoc faq, Tech. rep., ECHR (2021).
URL https://www.echr.coe.int/Documents/HUDOC_FAQ_ENG.pdf

[48] D. J. . J. H. Martin., Speech and language processing (3rd ed. draft) (2021).
URL https://web.stanford.edu/~jurafsky/slp3/

A Top 100 most common anchor texts for incomplete or erroneous references

[('NJB 2018/301', 133), ('RV 1991, 19', 64), ('NJ 1954, 378', 55), ('BNB 1995/243', 52), ('V-N 1997/733', 49), ('LJN AY8271', 47), ('NJ 2012/283', 42), ('NJ 1956, 157', 37), ('NJ 1964, 445', 34), ('BNB 1991/329', 32), ('ECLI:NL:CRVB:2011:BR1216', 32), ('RV 1997, 70', 31), ('NJO 1980, 7', 29), ('LJN AU4088', 29), ('RV 1990, 89', 29), ('BNB 1978/252', 27), ('V-N 1986, blz. 1125', 26), ('NJ 1952, 29', 25), ('NJ 1946, 323', 25), ('LJN AA3805', 24), ('V-N 2006/34.15', 24), ('NJO 1979, 1', 22), ('BNB 1997/365', 21), ('RV 1994, 89', 21), ('RV 2004, 92', 21), ('NJ 1956/157', 21), ('BNB 1991/142', 20), ('NJ 1961, 444', 20), ('BNB 2001/367', 20), ('V-N 2001/65.2', 20), ('RV 2006, 31', 20), ('BNB 2006/197', 19), ('NJB 1991, 50', 19), ('RV 2006, 43', 19), ('BNB 2006/252', 19), ('V-N 2009/29.17', 19), ('NJ 1961, 416', 18), ('LJN BJ2790', 18), ('LJN: BN0939', 18), ('ECLI:EU:C:2018:465', 18), ('BNB 2002/50', 17), ('ECLI:EU:2016:198', 17), ('BNB 1963/266A', 16), ('V-N 2001/11.27', 16), ('DD 97.301', 16), ('V-N 1997, blz. 653', 16), ('V-N 2003/10.25', 16), ('NJB 2004/17', 16), ('BNB 1979/314', 16), ('BNB 1992/313', 15), ('V-N 1973, nr. 18A', 15), ('NJ 1952, 314', 15), ('V-N 2001/14.2', 15), ('RV 1997/70', 15), ('NJ 1946, 712', 15), ('WFR 2008/6765', 15), ('Jur. 1987, 3696', 15), ('RSV 1960/67', 14), ('NJB 2000, p. 259', 14), ('NJ 1962, 266', 14), ('V-N 1969, blz. 880', 14), ('RV 1977, 87', 14), ('NJ 1962, 246', 14), ('LJN AT9173', 14), ('V-N 2009/2.7', 14), ('Jurispr. 1982, blz. 59', 14), ('BNB 2009/33', 14), ('BNB 1991/344', 13), ('NJ 1962, 89', 13), ('NJ 1955, 386', 13), ('NJ 1954, 242', 13), ('V-N 1996, blz. 1830', 13), ('Jurispr. 1979, p. 2188', 13), ('Jurispr. 1997, p. 7493', 13), ('NJ 1989, 449', 13), ('BNB 1990/341', 13), ('2', 13), ('TAR 2006, 8', 13), ('V-N 2004/23.6', 13), ('BNB 2008/140', 13), ('NS 2004/18', 13), ('ECLI:NL:RVS:2015:3064', 13), ('V-N 1996, blz. 1884', 12), ('NJ 2001, 721', 12), ('WFR 1997/6227', 12), ('M en R 1997/6', 12), ('NJ 1959, 102', 12), ('V-N 1997, blz. 1662', 12), ('2002', 12), ('V-N 1996, blz. 1545', 12), ('V-N 2004/37.7', 12), ('DD 1995, p.1056', 12), ('BNB 2013/45', 12), ('BNB 1978/7', 11), ('V-N 1994, blz. 3873', 11), ('RV 1992, 95', 11), ('BNB 2000/72', 11), ('NJ 1951, 356', 11), ('V-N 1997, blz. 1225', 11), ('V-N 2001/3.1', 11)]

B Top 100 most common anchor texts for ambiguous references

[('Cichowski', 2510), ('zich en anderen', 731), ('bono', 725), ('Muller', 664), ('Herrmann', 616), ('medici', 567), ('Sahin', 458), ('Dogan', 453), ('Salduz', 433), ('Hartman', 372), ('BNB 2005/337', 366), ('weeks', 294), ('Aksu', 164), ('de Afdeling (uitspraak van 6 maart 2008 in zaak nr. 200706839/1', 157), ('Liu', 153), ('Cunningham', 142), ('Ferrari', 137), ('Müller', 135), ('Dinç', 127), ('V-N 2003/52.2', 123), ('Coskun', 121), ('Bianchi', 105), ('Maksimovic', 89), ('BNB 2001/52', 87), ('Szabo', 80), ('Schwartz', 77), ('Ramsahai', 74), ('Karakaya', 72), ('Salah', 71), ('Thomson', 62), ('Soylu', 58), ('NJ 2013/241', 58), ('Bozbey', 56), ('Chamber of Commerce and Industry', 47), ('BNB 1990/61', 46), ('Lynch', 45), ('Knezevic', 45), ('arrest Paposhvili', 44), ('Demirtas', 43), ('Grimm', 42), ('Maurice', 42), ('Sheikh', 40), ('Simsek', 40), ('HR 17 december 2004, nr. C03/114HR', 40), ('BNB 1990/272', 39), ('BNB 2003/182', 38), ('HR 15 april 2011, nr. 09/03075', 37), ('HR 13 december 1989, nr. 25 077', 35), ('Aslan', 35), ('HR 29 november 2000, nr. 35 797', 34), ('BNB 1998/293', 33), ('de Afdeling (zie onder meer de uitspraak van 6 maart 2008 in zaak nr. 200706839/1', 33), ('BNB 2005/42', 32), ('Rey', 30), ('Ahmad', 30), ('Ersoy', 30), ('EHRM 14 mei 2013, nr. 66529/11', 30), ('Hoge Raad 29 november 2000, nr. 35 797', 29), ('Hoge Raad 15 april 2011, nr. 09/03075', 29), ('Kalmar', 28), ('EHRM van 11 januari 2007 in zaak nr. 1948/04', 28), ('Gümüs', 28), ('Morel', 27), ('Aydogan', 27), ('Sipos', 27), ('BNB 1991/339', 26), ('Mustafa', 25), ('sheikh', 25), ('NJ 2002, 385', 25), ('BNB 2000/306', 25), ('Hoge Raad 29 november 2000, nr. 35.797', 25), ('HR 29 november 2000, nr. 35.797', 25), ('Cetinkaya', 24), ('V-N 2006/42.13', 24), ('Güçlü', 23), ('HvJEU 16 juli 2009, nr. C-5/08', 23), ('Mann', 22), ('Hoge Raad van 4 december 1991, nr. 27 661', 22), ('HR 8 augustus 2003, nr. 36 769', 22), ('Ricci', 22), ('Hoge Raad van 15 april 2011, nr. 09/03075', 22), ('BNB 2005/ 337', 22), ('EHRM 3 mei 2001, no. 31827/96', 22), ('BNB 1995/15', 21), ('Neumann', 21), ('HR 18 april 2003, nr. 37790', 21), ('VN 2005/22.6', 21), ('Lewandowski', 21), ('HR 11 juni 2004, nr. 39 467', 21), ('HR 7 oktober 2005, nr. 35729', 21), ('BNB 1992/295', 20), ('BNB 1986/118', 20), ('Szegedi', 20), ('Zielinski', 20), ('Hoge Raad 8 januari 1986, nr. 23 034', 20), ('HR 8 augustus 2003, nr. 36766', 20), ('NJ 2003, 171', 20), ('Hoge Raad 18 april 2003, nr. 37 790', 20), ('Firat', 20), ('HR 18 april 2003, nr. 38 122', 20)]

C Top 100 most common anchor texts for all references

[('conclusie', 8611), ('Cichowski', 2907), ('vitale', 1813), ('Burg', 1282), ('boden', 1148), ('Salduz', 952), ('dagdelen', 915), ('Bosnak', 835), ('Jurgens', 814), ('Muller', 783), ('bono', 763), ('scala', 763), ('zich en anderen', 731), ('Haviltex-maatstaf', 730), ('BNB 2005/337', 728), ('Herrmann', 619), ('rees', 597), ('medici', 594), ('Haviltex', 585), ('Schreurs', 526), ('Sahin', 509), ('Dogan', 469), ('ECLI:NL:RVS:2016:3350', 461), ('God', 439), ('Haviltex-criterium', 426), ('dering', 419), ('Hartman', 412), ('ECLI:NL:CRVB:2009:BH1009', 384), ('LJN BH1009', 374), ('Bahaddar tegen Nederland', 368), ('', 366), ('ECLI:NL:HR:2010:BL1116, NJ 2011/237', 344), ('van der Sluijs', 339), ('ECLI:NL:HR:2014:3474', 329), ('beslissing over', 326), ('de Afdeling (uitspraak van 6 maart 2008 in zaak nr. 200706839/1', 314), ('ECLI:NL:HR:2016:252', 309), ('weeks', 299), ('ECLI:NL:CRVB:2012:BV1958', 295), ('Haviltexmaatstaf', 283), ('Okyay', 277), ('ECLI:NL:HR:2008:BD2578', 270), ('ECLI:NL:CRVB:2015:4920', 270), ('NJ 2000, 721', 264), ('de Afdeling van 4 mei 2005 in zaak nr. 200406320/1', 252), ('NJ 2004, 376', 251), ('V-N 2003/52.2', 246), ('dere', 245), ('Geffen', 243), ('JV 1998/45', 243), ('Salduz-jurisprudentie', 237), ('Unen', 236), ('Conclusie', 235), ('Haviltex-norm', 230), ('NJ 2008, 358', 227), ('ECLI:NL:RVS:2014:4129', 224), ('ECLI:NL:CRVB:2012:BV2512', 211), ('ECLI:NL:CRVB:2015:1550', 208), ('NJ 1996, 249', 202), ('ECLI:NL:CRVB:2015:1', 195), ('LJN BD2578', 186), ('ECLI:NL:CRVB:2015:3803', 185), ('ECLI:NL:CRVB:2016:4872', 184), ('NJ 1981, 635', 183), ('ECLI:NL:RVS:2014:188', 183), ('ECLI:NL:CRVB:2015:1663', 181), ('Müller', 175), ('Alcatel', 175), ('BNB 2001/52', 174), ('Römer', 173), ('Aksu', 173), ('CeH', 173), ('Haviltexnorm', 172), ('ECLI:NL:RVS:2016:2582', 171), ('Richard', 170), ('Ferrari', 168), ('BNB 2009/222', 168), ('Cunningham', 167), ('Albayrak', 165), ('ABU', 165), ('ECLI:NL:HR:2005:AO9006', 164), ('23 juni 2010 in zaak nr. 200908558/1/V6', 162), ('Davis', 160), ('ECLI:NL:HR:2011:BN6324', 160), ('NJ 1997/481', 156), ('Liu', 155), ('LJN: AR4716', 151), ('Abu', 149), ('NJ 2005, 493', 149), ('beslissing tot', 147), ('licence', 147), ('Bäcker', 146), ('ECLI:NL:HR:2010:BL1116', 146), ('ECLI:NL:CRVB:2014:3754', 145), ('ECLI:NL:CRVB:2017:2226', 144), ('van Vlimmeren', 143), ('LJN AY5142', 142), ('god', 141), ('Dinç', 141), ('Backer', 140)]

D Top 100 most common uri's for regular ECLI uri references (568,747)

[('ECLI:NL:HR:1981:AG4158', 3795), ('ECLI:CE:ECHR:1999:1102JUD003716697', 1961), ('ECLI:CE:ECHR:2003:0128DEC003476302', 1412), ('ECLI:NL:HR:2008:BD2578', 1235), ('ECLI:CE:ECHR:1987:1027JUD001093084', 1147), ('ECLI:NL:HR:2005:AO9006', 988), ('ECLI:CE:ECHR:2008:1125JUD000176703', 937), ('ECLI:NL:XX:2008:BH0402', 916), ('ECLI:NL:CRVB:2009:BH1009', 839), ('ECLI:CE:ECHR:2007:0828DEC001387605', 835), ('ECLI:CE:ECHR:2010:0112DEC002948107', 818), ('ECLI:EU:C:2001:401', 812), ('ECLI:NL:HR:2004:AM2533', 758), ('ECLI:NL:HR:2010:BL1116', 739), ('ECLI:CE:ECHR:1986:1017JUD000953281', 667), ('ECLI:CE:ECHR:2013:1008DEC003684704', 661), ('ECLI:NL:HR:2004:AO1427', 651), ('ECLI:NL:HR:2011:BO5046', 606), ('ECLI:CE:ECHR:2015:0414DEC007305813', 555), ('ECLI:NL:XX:1998:AG8817', 546), ('ECLI:NL:HR:2006:AU9130', 539), ('ECLI:NL:HR:2016:252', 532), ('ECLI:NL:HR:2014:3474', 512), ('ECLI:EU:C:2009:368', 511), ('ECLI:NL:RVS:2016:3350', 462), ('ECLI:NL:HR:1989:AB9521', 459), ('ECLI:CE:ECHR:2005:1103DEC003015502', 431), ('ECLI:NL:HR:2000:AA7309', 429), ('ECLI:NL:HR:2007:BA2802', 415), ('ECLI:CE:ECHR:1984:0522JUD000936281', 413), ('ECLI:CE:ECHR:2001:0918DEC004982599', 413), ('ECLI:NL:HR:2009:BH3079', 412), ('ECLI:CE:ECHR:2012:0710JUD002935306', 406), ('ECLI:NL:HR:2011:BN6324', 403), ('ECLI:NL:HR:2002:AD5163', 392), ('ECLI:CE:ECHR:2014:1104JUD002921712', 384), ('ECLI:CE:ECHR:2007:1204JUD006346800', 381), ('ECLI:NL:HR:2006:AZ0758', 380), ('ECLI:CE:ECHR:2013:0115DEC000448712', 378), ('ECLI:NL:HR:2008:BA8179', 377), ('ECLI:NL:HR:2008:BC4959', 375), ('ECLI:NL:CRVB:2012:BV1958', 359), ('ECLI:NL:HR:2010:BL2823', 355), ('ECLI:NL:CRVB:1995:ZB1495', 353), ('ECLI:NL:HR:1965:AB7079', 339), ('ECLI:CE:ECHR:2011:0121JUD003069609', 330), ('ECLI:NL:HR:2010:BJ9092', 323), ('ECLI:NL:HR:1995:ZD0328', 317), ('ECLI:EU:C:1995:31', 310), ('ECLI:NL:GHAMS:2007:AZ7033', 310), ('ECLI:NL:HR:2011:BP2975', 310), ('ECLI:NL:HR:2008:BC1962', 302), ('ECLI:NL:CRVB:2004:AR4716', 299), ('ECLI:CE:ECHR:2008:0205JUD004474604', 296), ('ECLI:CE:ECHR:2014:1021DEC004418312', 291), ('ECLI:NL:HR:2003:AE9049', 287), ('ECLI:NL:HR:2009:BI8771', 287), ('ECLI:EU:C:2009:716', 285), ('ECLI:NL:HR:1983:AG4575', 281), ('ECLI:NL:RVS:2014:4129', 274), ('ECLI:NL:RVS:2005:AT5088', 272), ('ECLI:NL:HR:2009:BH2815', 271), ('ECLI:NL:CRVB:2015:4920', 271), ('ECLI:NL:HR:1996:ZC2105', 270), ('ECLI:CE:ECHR:1999:0622DEC004675799', 268), ('ECLI:NL:CRVB:2006:AY9971', 268), ('ECLI:EU:C:2005:775', 263), ('ECLI:NL:HR:2004:AO5822', 262), ('ECLI:NL:HR:2014:1496', 262), ('ECLI:CE:ECHR:2015:0310DEC007184510', 261), ('ECLI:NL:HR:2007:AZ3178', 256), ('ECLI:CE:ECHR:2006:0330DEC001465302', 255), ('ECLI:CE:ECHR:1999:0518DEC002604994', 253), ('ECLI:EU:C:2004:236', 253), ('ECLI:NL:HR:1989:AD5725', 249), ('ECLI:NL:CRVB:2012:BV2512', 249), ('ECLI:CE:ECHR:2003:0610DEC002779395', 247), ('ECLI:NL:RVS:2014:188', 247), ('ECLI:EU:C:1999:534', 245), ('ECLI:CE:ECHR:2001:0802JUD005427300', 245), ('ECLI:NL:HR:2009:BJ7907', 242), ('ECLI:NL:HR:2009:BH1083', 242), ('ECLI:CE:ECHR:1998:0114DEC003183296', 241), ('ECLI:CE:ECHR:2005:0712JUD003622097', 240), ('ECLI:NL:HR:2011:BN3442', 235), ('ECLI:NL:HR:2002:AD5356', 232), ('ECLI:NL:HR:2011:BO5080', 232),

('ECLI:NL:HR:2007:AZ6638', 231), ('ECLI:NL:HR:2008:BC5012', 231),
('ECLI:NL:HR:2007:BA4909', 221), ('ECLI:CE:ECHR:1997:0409DEC002616595', 219),
('ECLI:CE:ECHR:2006:0329JUD006236100', 215), ('ECLI:NL:XX:1996:ZB6862', 211),
('ECLI:NL:HR:2008:BC2800', 211), ('ECLI:NL:CRVB:2015:1550', 209),
('ECLI:NL:CRVB:2009:BH1896', 208), ('ECLI:NL:RVS:2010:BM8823', 207),
('ECLI:NL:HR:2013:BY8101', 207), ('ECLI:NL:GHAMS:2009:BK4978', 206),
('ECLI:EU:C:2016:198', 206)]