MWCC: A Corpus of Malawi Criminal Cases

Amelia V. Taylor
ataylor@poly.ac.mw
University of Malawi, The Polytechnic and tNyasa Ltd, Data Labs
Blantyre, Malawi

ABSTRACT

We describe the creation of a corpus of criminal court judgments issued by the Malawian courts. We highlight opportunities and challenges in machine understanding of this text.

KEYWORDS

Legal corpus, Entity recognition, Text annotation and markup

ACM Reference Format:
Amelia V. Taylor. 2020. MWCC: A Corpus of Malawi Criminal Cases. In *Proceedings of the 2020 Natural Legal Language Processing (NLLP) Workshop, 24 August 2020, San Diego, US.* ACM, New York, NY, USA, 9 pages.

1 INTRODUCTION

This article presents the creation of a corpus of criminal case judgments issued by appellate courts in Malawi and our experiments in preparing this text to be used with machine learning algorithms.

In Malawi, legal researchers face significant challenges in accessing and searching for relevant information. The Malawi Judiciary Development program that ran over the years 2003-2008, found that "an inadequate provision of fundamental legal resources, such as books, case reports, statute books and gazettes, greatly constrains the performance of the judiciary in its administration of justice". In 2013, the Malawi Judiciary, with funding from the European Union introduced a case management system use in the High Court and the Director for Public Prosecution [6, 18]. This new system has improved the case registration process but suffers from bottlenecks in processes and document logging; few case documents and final judgments are stored on the system and most of these contain no meta-data [18].

In the last few years, MalawiLII[1] provides online access to some of the court judgments, laws and statues in Malawi [2]. MalawiLII does not support a system of citation that makes it possible to link statutory law, case law and secondary law or to search by "legal terms" and their specific interpretations.

In view of these challenges, we started the development of an automatic tool that (I) provides meta-data for criminal court judgments on MalawiLII by demarcating their text into components such as headers, introduction, body and conclusion and extracting meta-data such as names of judges, dates, court of hearing; and (II) that provides a useful classification of the judgments for legal research.

In this paper we describe the creation of the corpus used in these two tasks and our results regarding the first. The paper is structured as follows. In Section 2 we review relevant literature. In Section 3 we describe the steps we took in creating the Malawi Criminal Cases Corpus (MWCC) and discuss adding markup to the files. In Section 4 we describe the types of annotations of law and case citations we added to the corpus. In Section 5, by means of examples from our corpus we illustrate challenges in machine understanding of legal text. We conclude in Section 6 .

2 LITERATURE REVIEW

A corpus is 'a collection of examples of language in use that are selected and compiled in a principled way' [16]. A list of corpora containing legal text is given in [23, 26]. These vary in size and genre coverage [3]. A small number of the corpora listed specialise on criminal judgments. However, these are not available or maintained regularly and seem to have been developed to serve a specific research objective in mind only: the HOLJ House of Lords Judgments Corpus is a small, containing 188 texts, subset of the collection of the House of Lords Judgments and was used for summarisation and rhetorical structural annotation[14, 15]. The Corpus de Sentencias Penales 2005 - 2010 was used to study 'legal phraseology' [26].

There are also clusters of research around some corpora, e.g., corpora of Italian legal text have been used in generating dictionaries of legal terms [21], in analysing their usage [10], and to assist in translations [12]. Similarly, a corpus of Dutch legislation was annotated using the Metalex XML scheme [4] and then enhanced with meta-data regarding the document structure, external descriptive data and law citations meta-data [8]. Corpora of Geek Tax Legislation [19, 20] and Greek Supreme Court decisions [11] were enhanced with XML structural mark-up and annotations. These projects did not use machine learning but made use of linguistic features of the text, regular expressions or syntactic parsers and grammars for data extraction.

The process of machine understanding of legal text involves a great deal of semantic enhancement, in order to make explicit or machine understandable 'the flexibility, intuition and capabilities of the conceptual structures of the human languages' in readiness for Web 4.0 [5]. The reality is that most legal text that is being made available online at present is in an unstructured form, i.e.,

[1] malawilii.org
[2] Albeit not complete or up to date, this is an improved successor to the listing of judgments that were initially done on the SNDP webpage, which listed judgments, court documents, some of the laws and Constitution of Malawi http:\www.sndp.org.mw/index-archived.php

[3] Some contain a wide range of types of text, e.g., academic journals, textbooks, contracts, opinions, legislation, e.g., the American Law Corpus some contain only law reports but cover several types of cases from administrative to criminal cases, e.g., the Corpus of US Supreme Court Opinions, some contain legal text of historical importance for example, the Old Bailey corpus is a historical corpus covering 197,745 texts over 1674 - 1913, some are multi-lingual like the ones covering European legislation, e.g., JRC-Acquis, Bononia Legal Corpus
[4] http:\www.metalex.eu/

NLLP @ KDD 2020, August 24th, San Diego, US

has almost no meta-data, no annotations that makes it possible to hyperlink it and 'machine understandable'. Hence, current work is still largely focussed on taking a collection on *unstructured text* and adding markup and annotations. As we are dealing with written text, there is already an inherent organisation within the text itself. So in this sense, 'adding structure' means to extract from or externalise this organisation in the form of markup or annotations on the text. We included a discussion on this terminology in Appendix A. The degree of 'organisation' in legal text varies a lot, text containing legislation is said to be more organised than that of court judgments, and notarial contracts [4] are more organised than legislative texts, and among legislations, those referring to tax and administrative law [8] are more structured.

In the case of court judgments, there are differences in how courts within the same country and or courts in different countries structure the text. However, all court judgments share common elements. They all contain, usually in their introduction, information such as the courts of hearing, dates and case numbering or docket numbers, names of the judges and other legal parties involved. They all follow a certain legal rhetoric in which facts are presented, points of law are discussed and finally the judgment is concluded. There are also common conventions that are used for citing laws and other cases. Some of these regularities were used to develop and test algorithms that employ machine learning techniques to 'understand' legal text, e.g., resolution of names of legal parties such as judges [9], resolution of citations to laws or other cases [17], extracting citations to laws [25], automatic summarization of court judgments [7, 14].

The trend seems to be that researchers collect their own data and use that to develop or test algorithms; a particular data set may never be used in another study. Noting this trend, [22] sets out a best-practice guide for the collection and analysis of legal corpora for linguistic analysis to ensure a certain degree of generality of the research results found when using a custom corpora. Generalisation issues may come from the impact that the genre of the text within a corpus has for example on the task of assigning meaning to terms, e.g., collocates of "breach" across different corpora may belong to different definitions/ meaning of the word.

We think that there is a need to set our similar guidelines for the use of legal corpora for machine learning purposes. For machine learning algorithms, generalisation challenges can be even less obvious because of the interplay between the impact of the language models used, of the differences in size and type or 'genre' of the training data versus the test data. An experiment that measures text similarity of legal documents [27] showed that a word2vec model was better than a bag of words model and the size of the training data compared to that of the corpus impacts the accuracy of the similarity results. However, to explain these results the authors spend very little time describing their data apart from describing it as 'selected larceny cases'.

Putting together a corpus takes significant time and involves a diversity of linguistic and computing skills. In building the MWCC corpus we tried to ensure a certain 'separation of the corpus design' from research design in order to ensure that other researchers will use our corpus. We present the construction of a corpus of Malawi criminal case judgments from a set of 'unstructured' text files and describe some of our experiments in adding markup and annotations to the judgments. By means of examples from our corpus, we reflect on the importance of cooperation between linguistic and machine learning expertise in putting together legal corpora to solve challenges in machine understanding of the legal text. For reasons of space limitations we placed some important terminology definition in Appendix A.

3 THE MALAWI CRIMINAL CASES CORPUS (MWCC)

We followed the guidelines of [22] in creating the corpus.

3.1 The Target Domain for MWCC Corpus

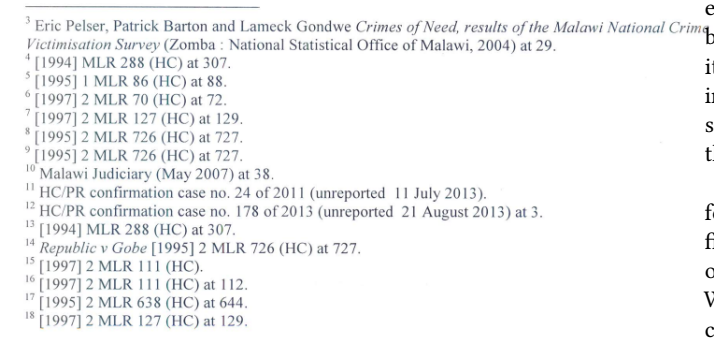
The data for MWCC corpus is the criminal case judgments stored in electronic doc, pdf and scanned images. These are obtained from the High Court Library. The librarian scans the physical judgments received from the High Court Registry, page by page, and stores them as pdf files. The physical papers are then catalogued in folders by year, and some of the scanned judgments are sent to law firms, judges and other parties which subscribe to this service. These are also uploaded to MalawiLII. The electronic scans have been named by the High Court Librarian according to a convention: [Case Name] [Case Type] [Case Number] [Case Year]. For example, *Lawrence Chibwana Vs The State Criminal Appeal No. 42 of 2010.pdf*. In some cases the name of the judge is also present in the title. In some cases, the naming of files does not correspond to their content, or names of parties have been misspelled.

The names of cases as retrieved from file names can be used to create a *case citator database* or if one exists to cross check them against that. To our knowledge the Malawi High Court Library does not maintain systematically a *case citator database*. For legal researchers, it is important to know which of these cases have been reported in official law reports as these receive a special naming convention. Identifying a citation is only useful if that can be 'resolved' and matched against an external knowledge source. A manual search for prior cases typically involves formulating a query (using party names, dates, docket numbers, and courts), retrieving documents from a database of millions of opinions, and iterating the process until the right cases are found. Challenges in case names resolutions were discussed in [17] where the authors describe the development of a tool that provide automated assistance to the citators of Thomson Legal and Regulatory. In some cases, a citation cannot be resolved if there is no sufficient data in the context or if the judge refers to case documents that are not available or numbered (e.g., references to affidavit documents attached to the case).

3.2 The Design and Collection of the MWCC Corpus

We collected 682 criminal court judgments issued over 2010-2019 by the High Court and the Supreme Court of Malawi. These were stored as scanned images of physical documents. The files were roughly organised on disk according to the year in which they were issued. The steps we took in the preparation of the text for the MWCC corpus are: (I) File cataloguing: re-name the files with shorter names,remove special symbols, and maintain a mapping

Figure 1: Example of footnotes in court judgments.


[3] Eric Pelser, Patrick Barton and Lameck Gondwe *Crimes of Need, results of the Malawi National Crime Victimisation Survey* (Zomba : National Statistical Office of Malawi, 2004) at 29.
[4] [1994] MLR 288 (HC) at 307.
[5] [1995] 1 MLR 86 (HC) at 88.
[6] [1997] 2 MLR 70 (HC) at 72.
[7] [1997] 2 MLR 127 (HC) at 129.
[8] [1995] 2 MLR 726 (HC) at 727.
[9] [1995] 2 MLR 726 (HC) at 727.
[10] Malawi Judiciary (May 2007) at 38.
[11] HC/PR confirmation case no. 24 of 2011 (unreported 11 July 2013).
[12] HC/PR confirmation case no. 178 of 2013 (unreported 21 August 2013) at 3.
[13] [1994] MLR 288 (HC) at 307.
[14] *Republic v Gobe* [1995] 2 MLR 726 (HC) at 727.
[15] [1997] 2 MLR 111 (HC).
[16] [1997] 2 MLR 111 (HC) at 112.
[17] [1995] 2 MLR 638 (HC) at 644.
[18] [1997] 2 MLR 127 (HC) at 129.

for the naming. There was also a need to correct misspellings of names of parties present in the title or remove duplicates of files. There were also cases in which several cases were scanned together and saved in the same one file, these we had to split. (II) Image adjustments: straighten, remove watermarks, remove imperfections due to the scanning process; (III) Batch OCR: Run page by page OCR obtaining text corresponding to each line (word by word) in the image, saving this in json files which also contain some text formatting information, such as distances between lines, and font sizes; (IV) Text Reconstruction and Corpus Creation: Reconstruct the text from the files obtained by OCR and create the corpus files in text and XML format. We used Python openCV to deal with watermarks and markings on the text; and we wrote a Python batch program to split and merge back the images, the ocr.space API [5] for the OCR on the images, then we used custom python code to process the json files returned by the OCR API.

The image preparation stage could be improved by using techniques for automatically detecting image features which, if known in advance, can be useful for improving the quality of the OCR: most judgments contain official stamps, some outside the text, some on top of the text, most contain signatures of the judges or official clerks. These can be isolated, or removed before the OCR.

The most tricky part of the OCR process on these judgments was the *presence of headers, footers and footnotes*. The headers usually contained pagination and/or name of the case contained in the document. The header could not be always removed automatically based on text features, such as font size or distance to the main body of the text, as in many cases the font was the same and the headers were too close to the main text of the judgment as to appear as a normal part of the text. The footnotes also cannot be removed automatically because they contain relevant legal information. The footnote example in Figure 1 contains several case citations, e.g., [1994] MLR 288 (HC) at 307. This is an incomplete citation where one part, the case name, is in the main judgment text and the case citation is in the footnote. The ocr.space API extracts all textual information including the footnotes but these are not distinguished from the rest of the text. Heuristics based on structural information such as indentation, differences in font sizes, distances from the main text, could be used to recognise footnotes with some success.

[5]https:ocr.space

Another challenge was the frequent use of quotations, where a judge was discussing points relevant to the case at hand using extracts from law or from relevant cases. Some quotations used block quotes or other quotation marks. Others used indentation, italics or syntactical clues by the use of specific keywords that indicate their presence. It may be beneficial to use extra processing steps (e.g., using Tesseract [6]) to identify the presence of quotes in the text and to mark these as special parts of the text flow.

The electronic files of the corpus are structured into folders, one for each year. Each judgment has three files corresponding to: text file for introduction, text file for body with each paragraph being on one line, TEI XML file with markup and judgment paragraphs. We also have a separate file that maps the names of each file in the corpus with the name of the raw data file.

3.3 MWCC Corpus Statistics

We can describe our corpus according to the criteria in [2] as a full text (each text in the corpus is unabridged), synchronic (covers the period 2010 - 2019 and hence there is not a 'noticeable' change over this period in the way language is used or any change in the vocabulary used), terminological (our text contains both general and specific legal terms), monolingual (but containing names of people, organisation, geographical places that are typical of Malawi). The corpus contains 1,572,956 tokens, 1,374,635 words (a word may appear more than once), 63,574 sentences and 22,124 paragraphs extracted from 682 documents. There are 29,238 unique words, with a lexical variation of 2.1%. Table 1 shows a breakdown of cases per top 10 judges and Table 2 shows the breakdown of cases per year and shows sizes of yearly sub-corpora.

We used Sketchengine [7] to analyse the corpus in terms of part of speech tags, word lists and collocations. Table 3 shows the main part of speech frequencies for words that appear at least 5 times and excluding non-words. These represent 80% of our corpus. The percentage distribution are calculated on the whole corpus. Nouns, verbs and prepositions appear quite frequently. We also notice frequent use of adjectives; here are the top fifteen adjectives: *criminal, other,low,such,first, guilty, same, unreported, reasonable, maximum, convict, public, present, appropriate, excessive*. Several of these are specific of the legal language. The top ten most frequent nouns are all specific to the legal language: *court, sentence, case, evidence, offence, appellant, section, person, court, theft*.

Using Sketchengine we could also analyse the language used in our corpus compared to other corpora. In particular, we can look at corpora built for the general English language use such as the English Web corpus 2013, an English corpus made up of texts collected from the Internet, containing 15 billion words. Compared to this corpus, in ours, we see a much heavier use of prepositions and a lesser use of verbs compared to nouns. We can also find those n-grams or multi-words which appear frequently in our corpus and very rarely in the comparison corpus, such as *criminal procedure, hard labour, maximum sentence, theft simpliciter, first offender, accused person, reasonable doubt*. Such a comparison can be used to extract features useful for a machine learning classification or topic extraction analysis.

[6]https:\github.com/tesseract-ocr/tesseract
[7]https:\sketchengine.eu

Table 1: Malawi Criminal Cases in MWCC by top 10 judges (out of a total of 35 judges) in order of number of judgments issued.]

Judge Name	No. Cases
CHIRWA, J. M.	106
KAMANGA-NYAKAUNDA, D.	65
KAMWAMBE, M.L.	71
KALEMBERA, S.A.	25
MADISE, D.T.K.	45
MBVUNDULA, R.	28
MWAUNGULU, D.F.	81
NYERENDA, K.	51
SIKWESE, R. S.	37
Percentage of Total (627/682)	92%

Table 2: Composition of the MWCC by year

Year	No. Cases	Tokens	No. Parag.	MAx. Avg. Parag. Len.
2010	85	162,960	2,096	232
2011	72	155,154	2,959	131
2012	20	54,149	720	189
2013	162	426,584	6,840	200
2014	85	141,115	2,066	96
2015	122	274,583	3,538	131
2016	46	106,069	1,273	128
2017	27	52,038	810	42
2018	42	153,572	1,454	157
2019	21	46,732	368	223
Total	682	1,572,956	22,124	232

Table 3: Main Parts-of-Speech (Items with frequencies higher than 5.) This represent 80% of our corpus.

Part of Speech	No. Items (Lemmas)	Freq.	Distribution
Noun	3,959	393,777	29%
Verb	1,247	231,126	17%
Adjectives	953	75,894	6%
Adverbs	448	58,793	4%
Prepositions	81	223,172	16%
Conjunctions	13	43,975	3%
Pronouns	29	53,546	4%
Numerals	27	7,702	1%

There is also language that is particular to certain judges, e.g., *theft simpliciter* is used mainly by judge D F MWAUNGULU.

While word-lists and lists of keywords give us some useful statistics about the composition of our corpus, they do not take into account the context in which terms occur. When looking at a specific sequence of tokens/ words, the context surrounding a keyword is important. Such an analysis is called concordance or collocation analysis. Concordances are useful in finding out relevant connections between words (modifiers of specific words) and also to reveal multi-words units, e.g., detecting names of organisations, High Court of Malawi, or names of legal functions such as court clerk, attorney to state council, etc.

Figure 2: Introduction Part for Judgment 1 of 2013

JUDICIARY
IN THE HIGH COURT OF MALAWI
PRINCIPAL REGISTRY
CONFIRMATION CASE NO 689 OF 2013
Being Criminal Case No. 719 of 2013 from the Second Grade Magistrate Court Sitting at Chikhwawa
THE REPUBLIC
Versus
MATEYU THOM
THE HONOURABLE JUSTICE KENYATTA NYIRENDA
Margaret Munthali, Senior State Advocate, for the State
Accused person, Absent/unrepresented
Mrs. D. Mtegha, Official Court Interpreter
ORDER IN CONFIRMATION

A collocation is a sequence or a combination of words that occur together more often that what would be expected by chance. The strength of collocation is measured by the LogDice score (the higher the code the higher the collocation). *Words Collocations* can help understand the usage pattern of key legal terms, e.g., top modifiers of *murder* as a verb are *brutally, mercilessly, allegedly*. These can indicate the seriousness of the crime and or the intention. The collocates of crime are *consequence, offender, alibi, criminal, circumstance* and the word 'criminal' has the strongest collocations with *dangerous, hardened, unknown, hardcore, habitual*. Collocates for key legal terms can be used in topic extraction and the classification of judgments.

3.4 Adding Structural Markup to the MWCC Corpus

We have two formats for the files of the corpus: (a) an all text format and (b) an XML TEI format [8]. All judgments contain a front cover with information on the parties, the court of hearing, the dates and number of the case, the coram who heard the case (includes the judge, attorneys and other judicial clerks). It is possible to automatically separate this part from the main body of the judgment. In the text only format of the corpus, we keep separate files for the introduction, an example is given in Figure 2, and separate files for the paragraphs of the body of the judgment, each paragraph is stored in one line of text.

We based our separation of the introduction from the rest of the body on algorithm that is (a) looking for the presence of specific terms such as *ORDER IN CONFIRMATION, RULING* and (b) using formatting differences such as distances between lines of text used in the introduction versus the rest of the text.

Subcorpus of Introductions We thus obtained a sub-corpus made up of only introduction parts of the judgments. Out of that, we created a dictionary of legal keywords from all introductions (Table

[8] https:\tei-c.org/

4 Appendix B) which were then used to extract the legal parties involved in a case: such as name of the parties, judge, etc. This external meta-data was then added into the XML version of the corpus as meta-data for each judgment. An example of this meta-data is given in Appendix B.

While our approach did not involve machine learning, there is scope to use our sub-corpus to test supervised learning approaches to extract this information. In [3], the authors did something similar to us in the sense that they extracted formatting features which were later used in a supervised algorithm for extracting headings from pdf documents.

3.5 Chunking and Proper Names Recognition

Chunking poses many challenges. Some judgments are very long and may contain long paragraphs. Table 2 gives an indication of the maximum average length of judgments per year: ranging from 90 to over 200 tokens. We debated whether to store the text line by line, to split it into sentences using punctuation or to group the text in the same logical paragraphs as they were in the original images. We opted for the latter. We wanted to make sure we capture situations in which entities of interest break across lines. For example, in some case citations, one line may contain the names of the parties and another line, the court and dates. We used a heuristic based on the distances between lines to re-arrange the text to match the original paragraphs. We did not use punctuation to split into sentences because the text contained many 'entities' or elements which make use of full-stops, e.g., numbers, references to sections of law.

We used the POS tagging for extracting parts of our text which was likely to contain references to laws and cases. The English TreeTagger PoS tagset used by Sketchengine struggled with proper nouns because legal text makes use of capitalisation of many words for legal terms such as laws, e.g., Penal Code, legal parties, e.g., Appellant, or legal functions, e.g., Court Interpreter, references to laws, e.g., "Section", or names or crimes, e.g., "Manslaughter". These were usually tagged as nouns, but at times they were tagged as proper nouns as in *I/PP thus/RB convicted/VVD the/DT accused/VVN of/IN the/DT offence/NN of/IN Manslaughter/NP contrary/NN to/IN Section/NP 208/CD of/IN the/DT Penal/NP Code/NP*, or even verbs as in *Whereas/IN MUSATOPE/NP CHAPOTERA/NP was/VBD charged/VVN with/IN the/DT offence/NN of/IN murder/NN of/IN Yohane/NP Makiyi/NP contrary/NN to/TO section/VV 209/CD of/IN the/DT Penal/NP Code/NP*. In this example, "section" is not capitalised, but it is tagged as a verb possibly because of the presence of 'to' which usually precedes the infinitive form of a verb. The shape NP-NP is the most common for 2-grams in our text, and may correspond for example to names of people or places, but also to legal terms such as *Appellant Andrew, Judge Mwase*, legal bodies such as, *High Court, or Detective Sergeant*, or *names of laws, e.g., Drugs Act*. It is therefore important to have a way of distinguishing these legal terms from the rest of the text to enable a more accurate tagging. Using a list of relevant legal keywords and their use in context, may help with improving the POS tagging for legal text. We hope to look at evaluating legal-specific POS tagging methods in a future research using the MWCC.

The names of people, places and organisations which are particular to Malawi are not easily recognised by existing language models. The names used in Malawi are of Bantu origin [24] with European influences, hence sometimes parts of names are recognised while others are not. Names of people frequently appear in our text. We will annotate our text with Bantu names of people and places. We think that the MWCC can be used for building a training set, of typical Bantu names to be used with recent advances in BERT and transformers. For example [1] used the BERT model to recognise names of entities in Bulgarian, Czech and Polish and in [13] BERT was used to recognise Chinese names.

4 ADDING ANNOTATIONS TO THE MWCC CORPUS

4.1 Law Citations

There are several types of reference to laws found in our text. For example, references containing only the name of the law/statue *The following offences involving dishonesty in the Penal Code are based on circumstances....* or *...the Control of Goods Act derives its procedure in criminal matters from the Criminal Procedures and Evidence Code.*

There are references containing labels and names of the law *Section 11 (2) of the Supreme Court of Appeal Act. or Section 283 of the Penal Code.*

There are more complex types such as references by means of anaphors spanning more than one line, or sentence, or paragraph. *Section 12 of the Act...*
section of the same constitution ...
...in the Penal code...theft from a person (section 282(a)); theft from a dwelling house (section 282 (b))..

Appendix C gives a more comprehensive list. We annotated each judgment with law citations: an example is given in Table 5 of Appendix B.

4.2 Case Citations

Case citations may refer to cases published in official law reports or to unpublished cases, each of these using different styles of citation. A citation from the Malawi Law Report is:

Republic v Chizumila and others [1994] MLR 288 (HC) at 307

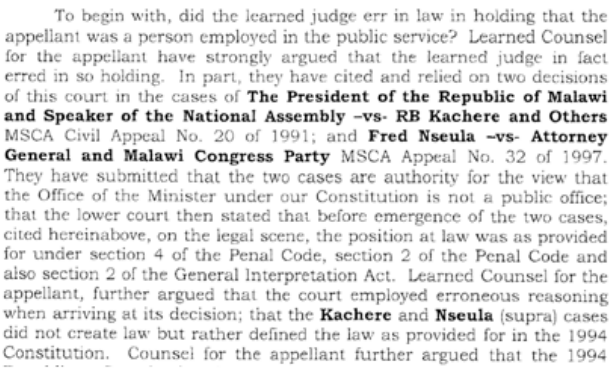
where Republic v Chizumila and others are the parties involved (also forming the case name), 1994 is the year of publication of the Malawi Law Reports, 288 is the case number and 307 is the location. Neutral citations were introduced in the UK in 2001 and are used by MalawiLII. For example, on MalawiLII the case:

Dalikeni and Others v The Republic (MSCA Criminal Appeal Case No. 6 of 2016)

is numbered as: *Dalikeni and Others v The Republic [2019] MWSC 8* where MWSC stands for Malawi Supreme Court and this is the eighth case registered on MalawiLII under this court. An example of unreported case is: *Republic vs Mpinganjira Bagala HC/PR confirmation case no. 24 of 2011 (unreported 11 July 2013)* where HC/PR stands for High Court Principal Registry.

The presence of names of people or organisation means that grammar rules or regular expressions cannot work on their own, and could be combined with lookup and some form of supervised learning. [25] used a supervised statistical models to extract standardised case citations of the type '[1994] MLR 288' from a selection

Figure 3: Example of Case Citation formatted in Bold - containing also a partial citation which needs resolution.


To begin with, did the learned judge err in law in holding that the appellant was a person employed in the public service? Learned Counsel for the appellant have strongly argued that the learned judge in fact erred in so holding. In part, they have cited and relied on two decisions of this court in the cases of **The President of the Republic of Malawi and Speaker of the National Assembly –vs- RB Kachere and Others** MSCA Civil Appeal No. 20 of 1991; and **Fred Nseula –vs- Attorney General and Malawi Congress Party** MSCA Appeal No. 32 of 1997. They have submitted that the two cases are authority for the view that the Office of the Minister under our Constitution is not a public office; that the lower court then stated that before emergence of the two cases, cited hereinabove, on the legal scene, the position at law was as provided for under section 4 of the Penal Code, section 2 of the Penal Code and also section 2 of the General Interpretation Act. Learned Counsel for the appellant, further argued that the court employed erroneous reasoning when arriving at its decision; that the **Kachere** and **Nseula** (supra) cases did not create law but rather defined the law as provided for in the 1994 Constitution. Counsel for the appellant further argued that the 1994

of 250 Pakistani court judgments. Their algorithms relied on training data in which case citations were manually tagged using the Inside-Out-Beginning notation. In a much larger project at Thomson Legal and Regulatory [17], a 'citator' database was available (containing a list of all available names of cases) and the task was to resolve the citations found into the citator. A Support Vector Machine (SVM) was used to improve the accuracy of the entity (name of cases) resolution. SVM were used also for entity resolution in [9] to match names of judge/attorneys and names of legal firms from text files with Westlaw records of attorney and legal firm files.

We think that, the extraction of case citations could, in some cases, be done directly from the scanned images, as most judges use italics of bold font when writing such citations. Then, a supervised algorithm that works on image data could be practical. However as shown in Figure 3, the convention used in the documents of our corpus is that only the case name is formatted differently not including the citation component. Some citations are partial, as 'Kachere and Nseula' shown in the image, and need to be resolved in context. In the next section we describe our experiments in extracting law citations.

5 EXPERIMENTS WITH SPACY

Our corpus served as an excellent data set to test extracting law and case citations and to generate test data for a supervised approach. SpaCy (https:\spacy.io/) is a Python library using state of the art neural networks for tagging, parsing and entity recognition. The Named Entity Recogniser in spaCy already has an entity for "LAW". For the English language, spaCy uses three models of varying sizes, small (sm), medium (md) and large (lg) trained using Convolutional Neural Networks on OneNotes 5.0 data set. The accuracy of the spaCy NER was reported to be over 80% for both precision and recall.

Our approach was as follows: we first used the standard spaCy NER to extract LAW entities, then we added an Entity Ruler to extract additional LAW entities. For example the pattern in Figure 4 of Appendix D matches references to sections which use two-level numbering, such as *Section 4 (a) or s. 4 (2) or section 42(2) (f)*. We used a Phrase Matcher based on a database of names of laws and statues in Malawi to extract LAWNAMES entities. We then merged these entities (e.g., the reference part merged with the law name) into larger ones and eliminated duplicates.

Most of the citations that are recognised by the standard SpaCy NER are of the type: *Section [number]*. However, SpaCy recognition depends on a uniform use of punctuation like spaces and full stops. So for example, if there are extra spaces, e.g, *Section 214 (a)* instead of *Section 214(a)*, the entity will not be always recognised. Also entities of the type *Sections 339 and 340* will also not be consistently recognised.

References to laws of England or laws that are typically found in other countries such as *Data Protection Act, Official Secrets Act* are recognised as these were present in the model. However, names of laws more particular to Malawi were not always recognised. Table 6 of Appendix D shows examples of law citations extracted using SpaCy and a comparison between the use of the lg vs sm SpaCy models: some entities which were found using the small model, sm, were lost when using lg, but overall, the use of larger model did result in a more accurate name identification of the law cited.

Table 7 of Appendix D shows the citations we were able to identify using in addition to the standard spaCy NER and then an enhanced method using both an Entity Ruler and a Phrase Matcher. The use of the Phrase Matcher allowed us to extract names of laws which are specific to Malawi. With this combination, we managed to find almost all the citations within the text. The phrase matcher was used to locate the complete names of laws referred to in the citations. For example, for the judgments of year 2010, spaCy NER managed to extract 507 valid citations (some incomplete). Using the enhanced process we extracted in total 1,162 which are citations (e.g., Section 224 A) and names of laws (e.g., Penal Code). When merged into full citations (e.g., Section 224 A of the Penal Code), we obtained a total of 611 citations. For the whole corpus, spaCy extracted 7,784 law citations out of a total of 18,929 obtained by the enhanced method. Overall, we extracted 10,390 law citations from our corpus. Thus, this process of extracting law citations worked reasonably well and can be used in constructing a training set of annotations for better results.

The case and law citations are stored in separate TEI files, an annotation file for each judgment file containing the paragraph, the exact position inside a paragraph, the text of the annotation and its type. The position of the annotations within a paragraph can also be used to resolve incomplete citations or anaphors. Some of the citations are incomplete and do not include the names of the law. For example the reference *section 235 (a)* appears several times in paragraphs 2 and 3, some occurrences do not contain the name of the law. The context of the judgment and the classification of the laws can help in the topic identification, e.g., section 235(a) of the Penal code covers issues of *causing grievous harm.*

6 CONCLUSION

We described the process of creating a corpus of criminal cases issued by Malawi courts. We reflected on the challenges and opportunities in semantically enhancing this text and the need for an intelligent pipeline that processes the text at all stages - some of the semantic enhancement can be done on raw images as we discussed for case citations. We would like to use our annotations and corpus for further training and classification.

REFERENCES

[1] Mikhail Arkhipov, Maria Trofimova, Yuri Kuratov, and Alexey Sorokin. 2019. Tuning Multilingual Transformers for Language-Specific Named Entity Recognition. Association for Computational Linguistics (ACL), 89–93. https://doi.org/10.18653/v1/w19-3712

[2] Atkins, Sue and Clear, Jeremy and Ostler, Nicholas. 1992. Corpus design criteria. *Literary and Linguistic Computing* 7, 1 (1992). https://doi.org/10.1093/llc/7.1.1

[3] Sahib Singh Budhiraja and Vijay Mago. 2020. A supervised learning approach for heading detection. *Expert Systems* (2020). https://doi.org/10.1111/exsy.12520

[4] María G. Buey, Angel Luis Garrido, Carlos Bobed, and Sergio Ilarri. 2016. The AIS project: Boosting information extraction from legal documents by using ontologies. In *ICAART 2016 - Proceedings of the 8th International Conference on Agents and Artificial Intelligence*, Vol. 2. https://doi.org/10.5220/0005757204380445

[5] Nuria Casellas. 2011. Semantic Enhancement of Legal Information… Are We Up for the Challenge? *VoxPopuLII* (2011).

[6] Winner Dominic Chawinga, Chaupe, Sellina Khumbo Kapondera, George Theodore Chipeta, Felix Majawa, and Chimango Nyasulu. 2020;. Towards e–judicial services in Malawi: Implications for justice delivery. 86:e12121 (2020;), 1–15. https://onlinelibrary.wiley.com/doi/epdf/10.1002/isd2.12121

[7] Min Yuh Day and Chao Yu Chen. 2018. Artificial intelligence for automatic text summarization. In *Proceedings - 2018 IEEE 19th International Conference on Information Reuse and Integration for Data Science, IRI 2018*. Institute of Electrical and Electronics Engineers Inc., 478–484. https://doi.org/10.1109/IRI.2018.00076

[8] Emile de Maat, Radboud Winkels, and Tom van Engers. 2006. Automated Detection of Reference Structures in Law. In *Legal Knowledge and Information Systems. Jurix 2006: The Nineteenth Annual Conference (Frontiers in Artificial Intelligence and Applications)*, Tom M van Engers (Ed.), Vol. 152. IOS Press, 41–50. http://www.leibnizcenter.org/docs/demaat/DeMaat-Jurix2006.pdf

[9] Christopher Dozier, Ravikumar Kondadadi, Marc Light, Arun Vachher, Sriharsha Veeramachaneni, and Ramdev Wudali. 2010. Named entity recognition and resolution in legal text. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, Vol. 6036 LNAI. https://doi.org/10.1007/978-3-642-12837-0_2

[10] R.R. Favretti, F. Tamburini, and E. Martelli. 2007. Words from Bononia Legal Corpus. *International Journal of Corpus Linguistics* 6, 1 (2007). https://doi.org/10.1075/ijcl.6.3.03ros

[11] John Garofalakis, Konstantinos Plessas, Athanasios Plessas, and Panoraia Spiliopoulou. 2019. Modelling Legal Documents for Their Exploitation as Open Data. In *Lecture Notes in Business Information Processing*, Vol. 353. https://doi.org/10.1007/978-3-030-20485-3_3

[12] Patrizia GIAMPIERI. 2019. the Bolc for Legal Translations: a Trial Lesson. *Comparative Legilinguistics* 39 (dec 2019), 21–46. https://doi.org/10.14746/cl.2019.39.2

[13] CHENG GONG, JIUYANG TANG, SHENGWEI ZHOU, ZEPENG HAO, and JUN WANG. 2019. Chinese Named Entity Recognition with Bert. *DEStech Transactions on Computer Science and Engineering* cisnrc (2019). https://doi.org/10.12783/dtcse/cisnrc2019/33299

[14] Claire Grover, Ben Hachey, and Ian Hughson. 2004. The HOLJ Corpus. Supporting Summarisation of Legal Texts. *COLING 2004 5th International Workshop on Linguistically Interpreted Corpora* (2004).

[15] Ben Hachey and Claire Grover. 2004. A rhetorical status classifier for legal text summarisation. In *In Proceedings of the ACL-2004 Text Summarization Branches Out Workshop*.

[16] Chu Ren Huang and Yao Yao. 2015. Corpus Linguistics. In *International Encyclopedia of the Social & Behavioral Sciences: Second Edition*. Elsevier Inc., 949–953. https://doi.org/10.1016/B978-0-08-097086-8.52004-2

[17] Peter Jackson, Khalid Al-Kofahi, Alex Tyrrell, and Arun Vachher. 2003. Information extraction from case law and retrieval of prior cases. In *Artificial Intelligence*, Vol. 150. https://doi.org/10.1016/S0004-3702(03)00106-1

[18] Binart Kachule and Amelia Taylor. 2018. Understanding the Factors affecting the Utilisation of the Case Management System of the Malawi Judiciary Conference: EGPA 2018, EGPA study group XVIII on justice and court administrationAt: Lausanne, Switzerland.

[19] Marios Koniaris, George Papastefanatos, and Ioannis Anagnostopoulos. 2018. Solon: A holistic approach for modelling, managing and mining legal sources. *Algorithms* 11, 12 (dec 2018). https://doi.org/10.3390/a11120196

[20] Marios Koniaris, George Papastefanatos, and Yannis Vassiliou. 2016. Towards automatic structuring and semantic indexing of legal documents. In *ACM International Conference Proceeding Series*. Association for Computing Machinery. https://doi.org/10.1145/3003733.3003801

[21] Paola Mariani and Costanza Badii. 2005. Methods and techniques for building a digital historic-law dictionary. In *Proceedings of the International Conference on Artificial Intelligence and Law*. 230–231. https://doi.org/10.1145/1165485.1165523

[22] James C Phillips and Jesse Egbert. 2017. Advancing Law and Corpus Linguistics: Importing Principles and Practices from Survey and Content-Analysis Methodologies to Improve Corpus Design and Analysis. *Brigham Young University Law Review* 2017, 6 (2017).

[23] Gianluca Pontrandolfo. 2012. Legal Corpora: an overview.

[24] Peter E. Raper. 2017. Indigenous common names and toponyms in Southern Africa. *Names* 65, 4 (2017), 194–203. https://doi.org/10.1080/00277738.2017.1369742

[25] Shahmin Sharafat, Zara Nasar, and Syed Waqar Jaffry. 2019. Data mining for smart legal systems. *Computers & Electrical Engineering* 78 (sep 2019), 328–342. https://doi.org/10.1016/J.COMPELECENG.2019.07.017

[26] Friedemann Vogel, Hanjo Hamann, and Isabelle Gauer. 2018. Computer-Assisted Legal Linguistics: Corpus Analysis as a New Tool for Legal Studies. *Law and Social Inquiry* 43, 4 (2018). https://doi.org/10.1111/lsi.12305

[27] Chunyu Xia, Tieke He, Wenlong Li, Zemin Qin, and Zhipeng Zou. 2019. Similarity Analysis of Law Documents Based on Word2vec. In *Proceedings - Companion of the 19th IEEE International Conference on Software Quality, Reliability and Security, QRS-C 2019*. https://doi.org/10.1109/QRS-C.2019.00072

A SOME DEFINITIONS

Markup. The markup adds what is usually called, *external information*, meaning information about the text. Legal markup for court judgments: case name, case number, court of hearing, date of case registration, date of judgment, judge, legal parties such as appellant and respondents, lawyers, court clerks.

Simple Structural Annotation. The word *structure* is used to mean a particular general arrangement that is present in most texts. The simplest arrangement can be one in which the text is arranged in paragraphs, or a text may be arranged in chapters or sections, or even more generally, as having the three main parts of introduction, a body and a conclusion. These structural components follow a tree-like hierarchy.

Complex Structural Annotation. In this sense, structure is dependent on the nature of the text. For example, a *case judgment* typically has portions of text in which the *facts* of the case are presented, followed by *proceedings* or the history of the case, e.g., previous rulings, *a discussion* of the relevant points of law and the *a conclusion* for the case. Structure may also mean *rhetorical styles* which are used in some part text.

Legal Annotations. The annotation in this case refers to locating specific pieces of text. This can be specific words, or phrases. Usually the pieces of interest appear next to each other in the text, but sometimes they do not. In the case of legal text, one is interested in (a) *legal terminology*; (b) *citations to laws and statues*; (c) *citations of other cases.*

Legal Resolution. Annotations with case citations or law citations need to be standardised so that documents can be hyperlinked.

Legal Classification. This usually refers to a semantic arrangement of the text into a predefined list of categories according to a pre-established criteria. For example, court judgments can be classified according to a court taxonomy, e.g., e.g., civil cases versus criminal cases vs. commercial cases. Some classification criteria are not linked to a taxonomy, e.g., one can classify court judgments based on the type of crime it mostly deals with say theft versus homicide.

Topic Extraction. Topic extraction attempts to discover the most important or relevant keywords in documents. so for example, one would use this to check if the text at hand contains health advice or a football match commentary. It is common to use topic extraction in order to classify documents.

(Un)Structured Legal Text Legal text is by nature quite well organise internally, however, by structured legal text we mean text that contains some or all of the above. Unstructured legal text are doc, pdf, scanned images of such documents that apart from being stored electronically, do not contain any of the above.

B CORPUS FILES EXAMPLES

```
<?xml version="1.0"?>
<TEI.2 lang="en" n="2010_17" id="judg_2010_17">
....
<titleStmt><title type="full">
<title type="main">Elizabeth Bonomali Vs The State</title>
<title type="sub">Criminal Appeal Case No 7 of 2010</title>
</title></titleStmt>
....
<catRef target="#courtofhearing">
<keywords>
<list type="courts">
<item>IN THE HIGH COURT OF MALAWI</item>
<item>PRINCIPAL REGISTRY</item>
</list>
</keywords>
....
<front>
<list type="caseinfo">
<item>CRIMINAL APPEAL CASE NO 7 OF 2010</item>
</list>
<list type="parties">
<item>ELIZABETH BONOMALI</item>
<item>THE REPUBLIC</item>
</list>
<list type="coram">
<item>HON JUSTICE J M CHIRWA</item>
<item>Mr Lemucha of Counsel for the State</item>
<item>Chipembere of Counsel for the Accused</item>
<item>N Nyirenda Official Interpreter</item>
</list>
</front>
<body>
<p n="2">The Appellant, Elizabethe Bonomali, was convicted
    after a full trial of the offence of unlawful wounding
    contrary to Section 214 (a) of the Penal Code and sentenced
     to 12 months' imprisonment with hard labour by the First
    Grade</p>
<p n="3"> Magistrate's court at Dalton Road, Limbe, on the 25th
    day of February, 2010. She has appealed to this Court
    against both the conviction and sentence.</p>
<p n="4">When the Appeal came up for hearing on the 26th day
    of March 2010 the Appellant indicated that she had
    abandoned her appeal against the conviction and that her
    complaint remained against the sentence only. I thus leave
    the conviction endorsed by the Learned Magistrate
    unfettered with.</p>
.....
</body>
```

Table 4: Keywords for extracting legal parties generated from the heading of judgments

Modifiers	Legal Functions	Case Parties
Chief	Reporter	Appellant
Senior	Advocate	Respondent
Principal	Interpreter	Applicant
Acting	Magistrate	Accused
Legal Aid	Justice	Defendant
Deputy	Prosecutor	State
Resident	Clerk	Convict
Principal	Recording Officer	Republic
Official	Judge	Plaintiff
Deputy	Lawyer	Coram
Court		Principal Witness
Honourable		Republic
Acting		Counsel

Table 5: Final Merged Entities for Judgment 17 of 2010 of MWCC

paragNo	Merged Entity	Start	End
2	Section 214 (a) of the Penal Code	117	150
5	Sections 339 and 340 of the Criminal Procedure and Evidence Code	897	961
7	Sections 339 and 340 of the Criminal Procedure and Evidence Code	1983	2047
8	Sections 339 and 340 of the Criminal Procedure and Evidence Code	88	152
10	Sections 339 and 340 of the Criminal Procedure and Evidence Code	183	247
12	Section 254 of the Penal Code	1034	1063
13	Sections 339 and 340 of the Criminal Procedure and Evidence Code	30	94
14	Section 339 (1):	0	16
15	section 283 of the Penal Code	477	506
15	Section 340 (1):	517	534
15	Section 339	792	815
16	sections 15 and 16	22	40
16	section 283 of the Penal Code	287	316
17	Section 339 of the	226	244
18	Section 340 of the Criminal Procedure and Evidence Code	68	123

C TYPES OF LAW CITATIONS

- References containing only the name of the law/statue
  The following offences involving dishonesty in the Penal Code are based on circumstances.... or *...the Control of Goods Act derives its procedure in criminal matters from the Criminal Procedures and Evidence Code...*
- References containing labels and names of the law
  Section 11 (2) of the Supreme Court of Appeal Act. or Section 283 of the Penal Code.
- References containing labels and abbreviations, or additional names in which a law is known (usually appears in brackets)
  section 6 of the Control of Goods (Import and Export)
  section 4 (d) of Part II of the Schedule to Bail (Guidelines) Act
  s. 149 of CP&EC
  section 17(d) and 42 of the Liquid Fuel and Gas (Production and Supply) Act
- References containing labels, names or abbreviations, and the year or date applicable to the law
  review of section 15 of the Code: it is commonplace that the CP&EC was amended in 2010
  section 340(3) of the Proceeds of Crime Act 2002 (POCA)
- References to laws that are pertaining to other countries (e.g., UK laws mentioned in Malawi court judgments)
  section 145 of the New Zealand Crimes Act of 1961
  offences against the Person Act, 1861 as held in R v Dica [2004] 2 Cr. App. R. 28
- references by means of anaphors spanning more than one line, or sentence, or paragraph.
  Section 12 of the Act...
  section of the same constitution ...
  ...in the Penal code...theft from a person (section 282(a)); theft from a dwelling house (section 282 (b))....
- References containing more than one label, number, e.g.,
  Section 2, 3 and 5 of ...

D RESULTS OF THE SPACY EXPERIMENTS

Figure 4: Example of pattern for extracting section citations for use with spaCy Entity Ruler

```
patterns = [{
"label":"SECLAW",
"pattern":[
  {"TEXT": {"REGEX": "^[Ss](ec\.?|ection|ections)$"}},
  {"IS_DIGIT":True,'OP':'?'},
  {"ORTH": "(",'OP':'?'},{},{"ORTH": ")",'OP':'?'},
  {"ORTH": "(",'OP':'?'},{},{"ORTH": ")",'OP':'?'},
  {"LOWER":"of",'OP':'?'}]
}]
```

Table 6: Example of improvements in precision but not recall using the lg versus the sm scaCy model.

Model	Parag	Pos. In Parag	Entity
sm	2	181	Penal Code
sm	46	86	section 187(1
lg	51	112	section 331
lg/sm	51	127	the Penal Code
lg	73	75	Bill of
lg/sm	82	33	section 328
sm	86	313	Act
sm	86	396	Act
lg	86	157	an Act of Parliament
lg	86	228	an Act of Parliament
lg/sm	86	29	Constitution
lg/sm	86	106	Constitution
lg	86	88	section 37
sm	86	376	section 4(1
lg/sm	86	320	the Official Secrets Act
lg	90	383	an Act of Parliament
lg/sm	93	115	Freedom of Information Act 2000
sm	93	151	the Data Protection Act
lg	93	151	the Data Protection Act 1998
lg/sm	95	42	Section 356

Table 7: Number of LAW Entities retrieved using the standard SpaCy model and by an enhanced method (+ EntityRuler and PhraseMatcher).

Year	SpaCy	Enhanced	Merged Entities	Spacy Recall
2010	507	1,162	611	44%
2011	554	1,310	635	42%
2012	153	400	184	38%
2013	3,406	8,432	4,769	40%
2014	621	1,640	863	38%
2015	1,044	2,414	1,378	43%
2016	469	1,055	589	44%
2017	236	616	295	38%
2018	597	1,374	772	43%
2019	197	526	294	37%
TOTAL	7,784	18,929	10,390	41%