EasyChair Preprint

№ 3506

Challenges in Machine Understanding of Legal Text

Amelia Taylor and Eva Mfutso-Bengo

May 30, 2020

Challenges in Machine Understanding of Legal Text

Amelia V. Taylor
Eva Mfutso-Bengo
ataylor@poly.ac.mw
ebengo@poly.ac.mw
University of Malawi, The Polytechnic
Blantyre, Malawi

ABSTRACT

The development of good models for representing legal text in order to make them suitable for machine-understanding and of models that incorporate human legal expertise into automatic tools, still pose great difficulties. In this research, we tackled the specific task of (a) creating a structured body of court judgments by annotating with key markup, legal citations and legal terms and (b) the problem of classifying court judgments according to the specific legal points. We document the creation of a corpus of Malawi criminal judgments (MWCC) and highlight opportunities and challenges in constructing a machine understanding of this text. We developed a pipeline which takes scanned images of criminal court judgments and creates structured documents in TEI format containing markups such as case name, case number, parties, coram and annotations of references to laws and other court cases which can be hyperlinked. We discuss the possibility of using these annotations and the International Classification for Crime Statistics to build an ontology for criminal cases useful for topic discovery and classification. The tools we used are Sketchengine, Spacy, Scikit-learn and Gensim.

KEYWORDS

Legal Text, ICCS, Case Metadata, Annotations, Case Citations, Law Citations, Classification of Crime, Corpus, Criminal Judgments, spaCy, Topic Classification

1 INTRODUCTION

Machine learning is seen as a key tool for machine understanding of texts and for uncovering hidden structures within them. Legal text is by nature fast growing. It is complex enough to provide excellent data for testing algorithms of data extraction, semantic analysis and machine learning. However, there are high quality requirements that are applicable to tools for legal text search and discovery before they can attract a strong user base. Advancing from syntactic to semantic tools remains challenging. Do machines learn from legal data in the same way as from other kinds of data? What kind of training data is needed to improve the efficiency of the machine learning algorithms for legal text? And how challenging is it to construct such data? We reflect on these questions by experimenting with a corpus of Malawi criminal court cases.

There is a growing body of legal text that is being made available on the internet by governments and other private and public organisations. The Legal Information Institute [1] has been publishing laws of several countries online free since 1992. The institute facilitates publications in countries where access to legal information is very poor, such as MalawiLII [2] for Malawi.

The legal system in Malawi is based on case law [15]. The decisions of the Supreme Court and of the High Court are binding precedents, while the decisions of the Magistrate Courts are of a persuasive nature. The law can only fulfil its appellate function when it is known and accessible. While knowledge in statues can be more easily retrieved as they are gazetted and organised on themes, knowledge on case law is more difficult to access.

MalawiLII is an 'open justice' effort which provides free of charge access to legislation and court judgments for Malawi. Despite the fact that it is not complete nor up-to date, it is a useful resource for legal researchers, considering for example that the official law reports in Malawi are issued with delays of years and there is an acute lack of legal commentaries. The information listed is unstructured, consisting of judgments alphabetically organised per year, judge and court of issue. The format is full text or scanned images of physical documents. The collection is not complete, documents contain no legal keywords, headnotes or summaries. MalawiLII also lists some of the legislation of Malawi although it does not have the rigour of paid platforms such as Westlaw [3] or Lexis Nexis [4]. On MalawiLII (and on other platforms on LII) search and retrieval is still largely a simple syntactic operation (e.g., using keywords that we hope are in the text). The current document structure 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. This linking of legal documents is crucial to legal research, particularly within a system based on case law such as that used in Malawi.

It is our desire to progress beyond the linking of legal documents by adding meta-data and annotations that enhance legal information search and knowledge management. For that reason, we use a combination of citations and their mapping to the International Classification of crime, to help classify the criminal court judgments of Malawi. We are interested in using available tools and testing their capabilities, in particular, using machine learning tools to organise the text and annotate it with formal structures (as opposed to keywords), for representing legal concepts for a better human-machine communication and understanding of legal text.

1.1 Challenges in Legal Research in Malawi

Law reporting is both part of and forms the basis for legal research. Doctrinal research is a two-part process in which the researcher

[1] https://www.law.cornell.edu/

[2] https://malawilii.org

[3] https://legalsolutions.thomsonreuters.co.uk/en/products-services/westlaw-uk.html

[4] https://www.lexisnexis.co.uk/

must first find and locate the sources of the law (usually covering several legal subjects) and then, interpret and analyse the text [11]. 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 "there is 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" and that "the situation is slowly improving in higher courts since the provision of internet access." On one hand, there are issues of accessibility and the availability of existing law (e.g., while some commentaries are found in law textbooks, law publications with commentaries and digests do not exist in Malawi) coupled with the scattered and untimely nature of the official reports [5].

On the other hand are the challenges coming from the fact that the current document structure of Malawi's legal text, e.g., court judgments, 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. There is no mechanism for linking case law and court decisions, e.g., no link between court judgments on MalawiLII and the Laws of Malawi (e.g., some are listed on MalawiLII). Malawian legal text (including court judgements) do not have the level of indexing that is used in such tools. For example, in the UK and the US, the West key number system and Shepard's Citations for Statutes provides a complete listing of each time a particular statute, regulation, or constitutional provision has been referred to and perhaps interpreted by a published decision of a court. A first subject index of (unreported) Civil and Criminal Cases 1997-2003 was compiled at the Judiciary by Heinrich Dzinyemba (Former High Court Librarian), print copy: it sorts by index and then gives bullet points that hint to the content of the case.

It is not clear how legal practitioners in Malawi make use of MalawiLII. It is quite likely, that judges and law firm maintain their own private libraries. Judges prefer to cite from hard copies for convenience and do not have access to online legal resources such as Lexis Nexis or Westlaw. Each judge uses their own format.

1.2 Approaches to Machine Representations of Legal Text

Law is more than datum in the sense that it is continuously made by the critical understanding and application of the law. Legal research includes a component of **content analysis**, i.e., a way of deconstructing the text, rather than synthesising meaning from text. Content analysis refers to reading the legal text, identifying categories, quantifying the use of words, examining the language, identifying patterns and themes within the data. This is used by legal scholars to identify meaning behind the words of the legal text. A doctrinal researcher, like a historian, is interested in the findings of legal principle applied by a judge by analysing case law, and by drawing logical conclusions about what the law is in that instance by analysing the legislation. In this research we are interested in content analysis in the sense that we annotate with citations to laws and cases (thus helping in the finding of the settled law that is applicable in each case) and we organise the text into categories/concepts and thus we identify differences between majority, preferred and better practice. In order to do this, our first task was to create a corpus containing the Malawi criminal cases.

1.3 Legal Corpora

A few legal corpora are available. The Bononia Legal Corpus [6] contains legislative, judicial and administrative documents in English and Italian, the JRC-Acquis Corpus [7] is a multilingual parallel corpus of European Union legislative texts, the Cambridge Legal English Corpus (not available online) contains books, journals and newspaper articles related to law. Corpora that are focussing on judgments issued by courts of law (especially in countries which use case law): the HOLJ consists of 188 judgments of the House of Lords 2001 - 2003, the British Law Reports Corpus [8] contains 1228 judgments issued by UK courts and tribunals between 2008 and 2010 full texts of which were obtained from BAILII [9]), the Corpus of US Case Law (CUSC) [10], containing published US court decisions digitised from the collection of the Harvard Law Library for the years 1760 – 1799, the Corpus of Supreme Court Opinions of the United States (COSCO-US) containing all opinions of the United States reports and opinions published by the Supreme Court through the 2017 term. Specific linguistic tools are available with the corpora such as KWIC concordances, word frequency lists, collocation statistics.

The corpora are usually put together by linguists for linguistic purposes, e.g., teaching the use of legal language to legal professionals. The users of these tend to be lexicographers, linguists including computational linguists. Computational linguists separate in two camps: those who rely and employ statistical methods to find and exploit regularities in the text and those who use the results of linguistic theory and logic as the foundation of language models. Computational linguists use corpora to test or train their models. For example, the HOLJ corpus was used for automatic summarization [10].

During our experiments for entity extraction using machine learning trained on general text models (which may or may not contain text of legal nature), we found that these do not perform well on court judgments, possibly due to the legal nature of the language used and its construction. There is a need to understand where weaknesses lie and how corpora can be used to train future models.

[5] The official Malawi Law Reports started in 1923 and the last volume was issued in 2014. There is work in progress for the missing years up to 2020, and the summary subject index covering the years 1990 - 2017 is work in progress with the legal publisher. The focus has been on the Supreme Court cases due to their importance and binding force and the selection of case is restricted to 60 per volume due to the publication space. These are selected by the Legal Research Unit (LRU) in conjunction with an editorial team which includes judges, academics and stakeholders such as the Malawi Law Society, the Law Commission, the Ministry of Justice, NGOs. The Legal Reporting Unit started an initiative to create summaries of judgments going forward, although logistically this is challenging; at the moment only one judge, Judge Sikwese, generates these summaries by analysing a print copy of a judgment using highlighters and margin notes that she types to add clarity and instructions for the typist later on. (This is an extract from our notes of a meeting with the Legal Reporting Unit of the Malawi High Court on the 21st January 2019).

[6] http://corpora.dslo.unibo.it/bolc_eng.html

[7] https://ec.europa.eu/jrc/en/language-technologies/jrc-acquis

[8] http://flax.nzdl.org/greenstone3/flax?a=fp&sa=collAbout&c=BlaRC&if=

[9] http://www.bailii.org/

[10] https://lawcorpus.byu.edu/cusc/concordances/search

2 THE MALAWI CRIMINAL CASES CORPUS (MWCC)

The data for MWCC corpus is the criminal case judgments stored in electronic doc files and scanned images[11]. The scans form the majority of the documents and we used an ocr pipeline to extract the text from these documents. The files have been named by the High Court Librarian according to a pattern: [Case Name] [Case Type] [Case Number] [Case Year]. For example, *Lawrence Chibwana Vs The State Criminal Appeal No. 42 of 2010.pdf*. Case number and year are part of the case citation. In some cases the name of the judge is also present in the title.

The names of cases as retrieved from the file names can be used to create a *citator database* or if one exists to cross check them against that. To our knowledge the Malawi High Court Library does not maintain a *citator database*. It is important to know which of these cases have been reported in official law reports as these receive a special naming convention. In some cases, the naming of files does not correspond to their content, or names of parties have been misspelled. When they appear as citations within the body of judgments, they frequently appear in an incomplete form or as implied citations. 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. The problem of matching against an external knowledge source was also discussed in [12] where the authors describe the development of a tool that provide automated assistance to the citators of Thomson Legal and Regulatory.

2.1 Processes involved in the creation of the corpus

We collected 682 criminal court judgments issued over 2010-2019 and saved as scanned images of physical documents. The files were roughly organised 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 naming: remove special symbols, re-name the files with shorter names and maintain a mapping for the naming; (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 structural 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 the desired format. We used Python openCV to deal with watermarks and markings on the text; and we used a Python batch program to split the images, the ocr.space API [12] 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. Another check which can automatically be done is to detect if there are more than one judgment in one file and to split these into separate files named appropriately.

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.

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 an indication of the case contained in the document, e.g., using the case name 'Rep vs. Banda'. The header could not be mechanically removed as in many cases it was 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 with some limited success.

Another challenge was the frequent use of quotations, where a judge is discussing points relevant to the case in hand using extracts from law or from relevant cases. Some quotations may be distinguished by the use of block quotes or other quotation marks. Others use 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 [13]) to identify the presence of quotes in the text.

2.2 The content of the MWCC Corpus

We can describe our corpus according to the criteria in [1] 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

[11] These are obtained from the High Court Library, which scans the physical judgments from the High Court Registry page by page, and stores them as pdf files. The physical judgments are then catalogued in folders by year, and some of the scanned judgments are sent to law firms, judges and other parties.
[12] https:ocr.space
[13] https://github.com/tesseract-ocr/tesseract

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
2010	85	162,960
2011	72	155,154
2012	20	54,149
2013	162	426,584
2014	85	141,115
2015	122	274,583
2016	46	106,069
2017	27	52,038
2018	42	153,572
2019	21	46,732
Total	682	1,572,956

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%. We used Sketchengine [14] to analyse the corpus in terms of part of speech tags, word lists and collocations. Collocations can help understand the usage pattern of key legal terms, e.g., top modifiers of *murder* as a verb are *brutally, mercilessly, allegedly*, and can be used in topic extraction and the classification of judgments.

We have two formats for the files of the corpus: (a) an all text format and (b) an XML TEI format [15]. 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 and separate files for the paragraphs of the body of the judgment. We generated keywords from all introductions (Table 3) which were then used to extract the legal parties involved

[14]https://sketchengine.eu
[15]https://tei-c.org/

Table 3: 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

in a case: such as name of the parties, judge, etc. This information is added as meta-data to the files, as can be seen in the example in Appendix A .

Our algorithm takes as input the output of the ocr, this is in the form of a json file containing text line by line, and for each line its component words, and generates as output the full text and the TEI files (an example can be see in Appendix A).

Chunking poses many challenges. Some judgments are very long and may contain very long paragraphs. We debated whether to store the text line by line 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, or citations span across more than one line. For example, one line may contain the case parties and another the court and dates. We used a heuristic based on the distances between lines to re-arrange the text in the original paragraphs in the document. We did not use the 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.

Tagging is also problematic. The English TreeTagger PoS tagset with Sketch Engine modifications struggles with proper nouns because legal text makes use of capitalisation of many words for legal terms succh as laws, e.g., Penal Code, legal parties, e.g., Appellant, or legal functions, e.g., Court Interpreter. Two grams of the shape NP-NP are the most common in the 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 more accurate tagging.

3 REFERENCE STRUCTURES IN LEGAL TEXT

3.1 Law Citations

There are several types of reference to laws found in our text:

- References containing only the name of the law/statue
  The following offences involving dishonesty in the Penal Code are based on circumstances....
  ...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.
  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 ...

Similar types of references found in Dutch Tax and Customs Administration text were described in [8]; the authors used a parser with hundreds of grammar rule for capturing the multiple types of formats found in law citations.

3.2 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. For example, a citation of a case from the African Law reports Malawi series is

McCarthy v. Stafford, 1923-60 ALR Mal. 4. [92]

where 'McCarthy v. Stafford' is the name of the case, '1923-60' is the year of the publication, the publication name is 'ALR Mal.', volume 4 and location 92. 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. Neutral citations are independent of the printed series of reports, instead the abbreviation used stands for the court of hearing and the number indicates the case number: [Year] Court Abbreviation [The number of the case]. The number of the case is different than the number used by the court. For example, on MalawiLII the case:

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

becomes

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. Some citations for unreported cases in the MWCC are of a form that resembles somehow the neutral citation:

Republic vs Kotamu (2012) Confirmation Case No. 180 (unreported).

In [12] a distinction is made between *direct history* which refers to citations within the same appellate chain and *treatment history* which refers to citations of other relevant cases. The retrieval of treatment history is done using thousands of grammar rules and an iterative construction of *pragmatic frames* in order to construct the context in which treatment history citations appear; these frames are based on the characteristic language judges use to introduce arguments and supporting facts from other cases and on matching into the citator database. Support Vector Machine (SVM) are 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. [18] used statistical models for extracting law and case citations from a set of 250 Pakistani civil proceedings and reported high levels of precision (over 80%) and recall (over 70%) for some of the experiments. From the examples given, it seems that the citations are those of reported cases in official law reports which display a uniform format (e.g., most citations in the footnotes shown in Figure 1).

In the next section we describe our experiments in extracting law citations. A similar process was used for extracting case citations.

4 EXPERIMENTS WITH SPACY

SpaCy [16] is a Python library using state of the art neural networks for tagging, parsing and entity recognition. A particular attraction for us was its flexible API and the extendability of its models. It also has a nice library for visual display. The Named Entity Recogniser in spaCy already has an entity for "LAW". For English, 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 [17]. The accuracy of the spaCy NER [18] was reported to be over 80% for both precision and recall. We found lower accuracy numbers of spaCy NER for extracting the entity LAW.

The *entity ruler* is designed to integrate with spaCy's existing statistical models and enhance the named entity recognizer by using pattern recognition. Each entity label is associated with a pattern.

[16] https://spacy.io/
[17] https://catalog.ldc.upenn.edu/LDC2013T19
[18] https://spacy.io/models/en

Figure 2: 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':'?'}]
}]
```

For example the pattern in Figure 2 matches references to sections which use two-level numbering, such as *Section 4 (a) or s. 4 (2) or section 42(2) (f).*

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. 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 (the reference part with the law name) into larger ones and eliminated duplicates.

Example of merging entities Consider the following paragraph: *This matter is before this court for review under section 42(2) (f) (viii) of the Constitution of the Republic of Malawi, under section 25 and 26 of the Courts Act, and under section 360 and 361 of the Criminal Procedure and Evidence Code.* Our entity extraction process identifies the following, where the two numbers represent the start and the end of the citation within the text:

section 42(2) (f), 50, 67
the Constitution of the Republic of Malawi, 78, 120
section 25 and 26 of, 128, 148
the Courts Act, 149, 163
section 360 and 361 of the Criminal Procedure and Evidence Code, 175, 238

The first two entities can be safely merged into one continuous citation. The same holds for the second and the third citation.

Most of the citations that are recognised by the standard SpaCy NER are of the type: *Section [number].* References to laws of England or laws that are typically found in other countries such as *Data Protection Act, Official Secrets Act* are also recognised. The use of the Phrase Matcher allowed us to extract names of laws which are specific to Malawi.

Table 4 shows some examples of law citations. It also shows in comparison that the use of the larger model lg does not lead to an improvement, as some entities which were found using the small model, sm, were lost. The use of larger model did result in a more accurate name identification of the law cited.

Table 5 shows that the recall rates for the standard spaCy NER pipeline does not exceed 45% when compared to the total number of citations we were able to identify using in addition to the standard spaCy NER, an Entity Ruler and Phrase Matcher. Thus, 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. These are not citations that may appear more than once in the corpus.

Table 4: 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

This process of extracting law citations works reasonably well and can be used in constructing a training set of annotations for better results. The position of the annotations within a paragraph can also be used to resolve incomplete citations or anaphors.

The case and law citations are stored in separate TEI files, each annotation specifies the judgment file, the paragraph, the exact position inside a paragraph, the text of the annotation and its type. We would like to use these annotations for a process of training and classification. We explain our strategy in the next section. As a means of justifying our approach we discuss challenges in using topic extraction on legal text.

4.1 Topic extraction using Gensim LDA

This section describes the challenge of running the Latent Dirichlet Allocation (LDA) using Gensim [19] on our corpus. In the LDA model [4], documents are represented by a mixture of topics which are characterised by a distribution over words. Words have numerical representations given by their Term Frequency-Inverse Document Frequency value to represent their significance in the document. Each word in the document is 'generated' by a single topic. We

[19]https://radimrehurek.com/gensim/models/ldamodel.html

Table 5: Number of LAW Entities retrieved using the standard SpaCy model and by an enhanced method.

Year	SpaCy NER	+ EntityRuler and PhraseMatcher	Merged Entities	Spacy NER 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%

Figure 3: LDA (run with 6 topics, 6 words per topic) on Judgments of 2019 from MWCC Corpus (extract)

Topic 0

0.408*"court" + 0.332*"accus" + 0.302*"evid" + 0.237*"case" + 0.216*"person" + 0.154*"section"

Topic 2

-0.414*"sentenc" + 0.309*"evid" + -0.281*"convict" + -0.179*"death" + -0.166*"circumst" + -0.150*"court"

Topic 3

-0.232*"sentenc" + 0.223*"time" + -0.213*"convict" + 0.206*"bail" + -0.173*"case" + 0.169*"procedur"

did no training, and pre-processing is simple (using standard stopwords and the Porter Stemmer for tokenization).

Figure 3 shows three of the topics identified in the subcorpus of 2019 judgments with Gensim LDA run with 6 topics and 6 words per topics (coherence score was 0.385). Similar results were obtained for other subcorpora and for the whole corpus.

These topics are as expected as they naturally follow from groupings of the statistical significant of keywords in the corpus. Here are the top 15 most significant terms: *court, case, evidence, person, section, offence, code, state, day, sentence, trial, appellant, prosecution, law, appeal.*

Topic 0 is not useful as it tells us nothing at all about the topic of a case. Topic 3 is slightly more informative as it indicates a bail case. Topic 3 may be seen as indicative of homicide because of the presence of the words murder, sentence and convict.

While such topics may be useful when looking at a diverse corpus containing several topics that are quite far apart from each other (say topics characteristic of civil cases as opposed to topics characteristic of criminal cases), in order to uncover useful topics within criminal judgments, we need to make use of legal insights into the text.

Example Section 209 of the Penal Code is indicative of murder/ intentional homicide. Two of the 2019 cases: *The State v Hanwell Ng'ambi, Owen Mtawali, Murder case no. 171 of 2018 01* and *The Repbulic vs Maxwell Matchina Sosola and 11 others; Homicide case No. 13 of 2018* contain references to Section 209 of the Penal Code. These were manually classified by a legal expert as dealing with 'intentional homicide'. The latter case is more complex containing several legal issues (e.g., murder, harm to a person with disability, transacting in human tissue, extracting human tissue, possession of human tissue). These may not appear in the text with these exact keywords but by means of law citations which themselves contain these keywords or related ones. In this case, these are Section 224B (1) of the Penal Code, Section 224 A (a) (i) of the Penal Code, Section 224 A (b) (ii) of the Penal Code, Section 224 (A) (e) of the Penal Code, Section 14 (1) Trafficking in Persons Act. All these citations are relevant to a meaningful classification of the case and for topic extraction.

Appendix B shows the list of citations extracted from the 10th judgment of 2019 in our corpus. 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.*

5 USING THE ICCS FOR TOPIC EXTRACTION

In this project we extended the use of International Classification for Crime Statistics (ICCS) [3] to attach meta-data to court decisions. The ICCS is applicable to all forms of crime data, whatever the stage of the criminal justice process (police, prosecution, conviction, imprisonment) at which they are collected. This approach opens future opportunities to relate data across the justice system [21] which could not be achieved with indices used in Malawi Law Reports (MLR) or MalawiLii. The recently introduced Case Management System at High Courts [6, 13] is not publicly accessible so that additional law reporting is required for access to legal information. The existing keyword function is underutilized and meta-data is lost when printed judgments are scanned before being sent to subscribers [19] . The Crime statistics of the Malawi police use 11 categories and, in 2019, 43% of the reported crimes were categorized as "other" [16].

ICCS has four levels of classification. The number of digits of the code number increases with each level. Most ICCS Level 1 6

Table 6: ICCS Level 1 Categories

Section	Description
1	Acts leading to death or intending to cause death
2	Acts leading to harm or intending to cause harm to the person
3	Injurious acts of a sexual nature
4	Acts against property involving violence or threat against a person
5	Acts against property only
6	Acts involving controlled psychoactive substances or other drugs
7	Acts involving fraud, deception or corruption
8	Acts against public order, authority and provisions of the State
9	Acts against public safety and state security
10	Acts against the natural environment
11	Other criminal acts not elsewhere classified

categories are based on the *telos* of the crime, in form of "Rechtsgüterschutz" or nature of protected legal good (against persons, property or both, life, public order, state security, environment) with exceptions of Sections 06 (involving drugs and controlled substances) and 07 (fraud, deception or corruption) which focus on activity and 11 (other crimes, which includes crimes under universal jurisdiction as the most important sub-group). Level 2 classification tend to be based on the activity, while level 3 and 4 classifications tend to be differentiations based on object of the crime (e.g. public, personal or business property or the victim being adult or child).

ICCS is relatively new and efforts to incorporate it in several countries are underway [20]. In [22] crimes were distinguished based on their severity (e.g., high vs low severity); classes of severity corresponded to groupings of the sections of the Level 1 of the ICCS. The authors used implementations of fuzzy fingerprint (FFP), Naïve Bayes (NB) and Support Vector Machines (SVM) - with the first being the most accurate of the three. Apart from this classification along severity lines [22] we are not aware of other attempts to use ICCS for machine classification.

Unlike victim or witness statements, or police reports, court judgments, which form the corpus of our study, include a legal assessment of the criminal incident. Judges observe good legal practice with a heading to the nature of their decision and, quite frequently, with an introductory paragraph which includes the core section and/or name of the offense in the charge. For example: *The Appellant in this matter Lawrence Chibwana was convicted of the offence of bringing in property dishonestly acquired outside the country contrary to section 331, Penal Code.* (File 1, 2010, of MWCC Corpus, paragraphs 2-4.)

We prefer the term "core offense or core section", which is the focus of the charge and decision, because a judgment can apply or interpret other key sections later in the text e.g. on general principles of criminal law or defenses that can be applied on all types of offenses. Only in exceptional cases this introductory paragraph is preceded by extraordinary case history or remarks. At this stage, the criminal process the legal syllogism of subsumption in which the Is-world is related to the Ought-world, which Hans Kelsen (1881-1973) analysed in his ground-breaking work [14] was carried out by the judge and/or prosecutor. This gives a leeway for a section-based ICCS classification for judgments.

The authors manually mapped section 38-409 (=Part 2) of the Malawi Penal Code Cap 7:01 according to Level 1 (11 items) and Level 2 ICCS sub-levels (63 items). Some ICCS categories are addressed in additional pieces of criminal legislation (Financial Crimes Act 2017, Trafficking in Persons Act Cap 7:06, ...). Due to the reductionist effect of law, that sees the complexities of the world through a lens of legal relevancy (for our purposes the elements of *actus reus* and *mens rea*), ICCS Level 3 and 4 with their additional disaggregation tags, would, in majority of cases, require additional information. This information may be found in the passages about the facts of the case. A first instance judgment or case docket and cannot be solved by referring to a statute only, that describes the offence in general terms. *Hence a machine classification of the judgments will need to use both legal meta-data and the mapping of the law into the ICCS.*

We wish to stress, that the subsumption of a criminal incident under a section of the Penal Code does not allow a conclusion of guilt, since evaluation of evidence, defenses and exculpation are ignored. A thematic classification uses an approximate subsumption under the most characteristic elements of the crime which is sufficient for purposes of content tagging for doctrinal legal research. ICCS itself provides inclusion descriptors in form of definitions that describe the most characteristic elements of the offence, list of specific examples with names of offences, results or modes of commission, cross-references to headings of sublevels and their descriptors. ICCS can strengthen the growing body of legal thesaurus and ontologies of law [2, 5, 7, 17] and criminal law in particular with regional or national mapping as additional "arms and legs". The ICCS does not differentiate linguistically in its inclusion between synonyms, near-synonyms or hyponyms except when referring to sublevels. Exclusion criteria are not antonyms. Exclusions rather serve as a demarcation for single classification in overlaps or can reduce the scope of any hypernyms or near-synonyms used in inclusions. We need to be aware that statistical subsidiarity for counting and legal subsidiarity for sentencing may differ.

In the mapping process narrow or minor concepts are subsumed under wider or major concepts, the offence in its legal abstraction under the ICCS category. The techniques employed are the matching of words, synonyms, notions, the comparison of elements of the crime in the Penal Code and elements in the ICCS description.

5.1 Challenges in Mapping into ICCS

We identified the following constellations. Firstly, there are split or partial matches which can occur for simple structural reasons because one section includes several types of offences in the subsections or variations within a clause (each of which would be a separate criminal incident which can be matched exactly.) This can be solved by an exact citation of the law and choosing subsections or variations as smaller units for the correspondence chart. The classification of judgments will then depend on the accuracy of the

[20] https://www.unodc.org/unodc/en/data-and-analysis/statistics/iccs.html

judge in citation whether referral to the facts of case is required not.

Secondly there are partial matches of sections, because the Malawian notion of the offence is wider than the ICCS category e.g. mercy-killing would amount in the Malawi legal system to murder, while ICCS differentiates between intentional homicide (which includes murder as hyponym) and euthanasia. Such incongruence is typical in comparative law discourses. An exact solution for classification of a judgment would analyse the facts of the case. If a single match for a section was required – which could translate into an error when mapping the judgment - the crime of greater likelihood would be the appropriate choice for approximation. The criminal incident itself would be a single match from the ICCS perspective, while the section in the Penal Code would lead to two matches.

Thirdly, there are borderline cases that are only prima facie double matches of sections (or criminal incidents) but which can be turned into a single-match by applying exclusion rules in the ICCS, e.g. violence used in a robbery does not count as a separate assault.

Fourthly, there are real double matches or borderline cases for which the ICCS does not provide a rule of demarcation or subsidiarity, e.g. Section 61 Penal Code Cap 7:01 (Malawi) "defamation of foreign princes" violates the honour of a foreign dignitary but extends beyond defamation as it aims to protect the security of the state by avoiding provocations, because it is part of Chapter VIII "Offences affecting relations with foreign states and external tranquillity." which would relate to ICCS Section 09 affecting state security. Legal reasoning would employ an analogy on how other acts like assault against a person in ICCS Section 02 relate to ICCS Section 09 or transfer ideas of "the more specific shall prevail over the more general norm in case of a conflict of norms" or of prioritization according to degree of gravity into this context. However, another point of view could postulate that Level 1 ICCS refers only to the direct object of the crime and not the *telos* for better feasibility and ease at data collection or use of AI.

All these constellations are exceptions and the ICCS in total offers a comprehensive and valuable classification system. We also need to clarify that a judgment can and should count for two ICCS classifications due to procedural technicalities when several independent criminal incidents are combined in one trial which includes several counts of charges. The counting units in ICCS are acts that constitute the criminal offence (p.11 of [21]). When two offences are committed simultaneously, e.g. discharging of a pollutant in the knowledge that it will kill fish and humans, it will depend on national charging or statistical counting rules whether both "intentional homicide" and "acts that cause environmental pollution" are captured or only the category in which the most serious offence was committed (principal offence rule) – a shortfall which ICCS is aware of (p.106 of [21]). We hope it will be clarified in the currently developed ICCS implementation manual [20]. This affects the comparability of data across jurisdictions.

For judgment classification we followed the Malawian charging pattern. Statistics and informatics inherit - but can also benefit from - underlying legal problems and controversies about the criminal act/"Tatbegriff" and rules of subsidiarity, and the concept of protected legal good "Rechtsgüterschutz" as justification for criminalisation which can coincide or extend beyond the directly protected object in the norm. In Malawi section-based correspondence charts can facilitate a classification up to level 2 with a few double matches but exact single-matching would in those cases require recourse to the full text of the judgment. If there is less emphasis on legal reasoning in a judgment, a double classification would even increase chances for information retrieval; statisticians and criminologists are interested in numbers of cases or criminal incidents that are comparable across a legal system.

Reduced significance of key words for classification. Another observation is that seemingly specific terms are used for different categories, though manual mapping can identify the different context e.g. prostitution falls in three different categories depending on context; working as prostitute (ICCS Section 08 Acts against public order, since it is considered as consensual sex), benefiting from another person's prostitution (ICCS Section 03 Injurious acts of sexual nature) and trafficking in human persons for prostitution (ICCS Section 02 Acts intending harm or intending to cause harm to the person) which is consistent with the underlying logic of ICCS. Similarly the word "fraud" in the context of deceit is diluted since the Penal Code uses the word "fraudulently takes" in the context of theft. (Section 271 Penal Code Cap 7:01 (Malawi)). Key words give direction but loose significance without context.

6 CONCLUSIONS

We presented the creation of a corpus of criminal cases. This can be useful in training and testing models for extracting law and case citations. We discussed challenges in using topic extraction algorithms given the complexities of legal text. We discussed the use of the ICCS in classifying and organising judgments according to topics. We believe that machine learning has a role in supporting legal research and we see our work as a contribution by means of a corpus and insights towards the development of a better machine understanding of legal text.

ACKNOWLEDGMENTS

This work has been funded by Artificial Intelligence 4 Development (AI4D) programme as part of the 1st Call for Project Proposals on Artificial Intelligence 4 Development Technologies.

REFERENCES

[1] 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

[2] Maria Angela Biasiotti and Daniela Tiscornia. 2011. Legal Ontologies: The Linguistic Perspective. In *Approaches to legal ontologies*, Giovanni Sartor (Ed.). Springer, Dordrecht, 143–166. https://doi.org/10.1007/978-94-007-0120-5{_}9

[3] Enrico Bisogno, Jenna Dawson-Faber, and Michael Jandl. 2015. The International Classification of Crime for Statistical Purposes: A new instrument to improve comparative criminological research. *European Journal of Criminology* 12, 5 (2015), 535–550. https://doi.org/10.1177/1477370815600609

[4] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent Dirichlet allocation. *Journal of Machine Learning Research* 3, 4-5 (2003). https://doi.org/10.1016/b978-0-12-411519-4.00006-9

[5] Pompeu Casanovas, Monica Palmirani, Silvio Peroni, Tom van Engers, and Fabio Vitali. 2016. Semantic Web for the Legal Domain: The next step. *Semantic web* 7, 3 (2016), 213–227. https://content.iospress.com/articles/semantic-web/sw224

[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] Vytautas Čyras, Friedrich Lachmayer, and Kristina Lapin. 2015. Structural Legal Visualization. *Informatica* 26, 2 (2015), 199–219. https://doi.org/10.15388/Informatica.2015.45

[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] 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).

[11] Terry Hutchinson and Nigel Duncan. 2012. Defining and Describing What We Do: Doctrinal Legal Research. *Deakin Law Review* 17, 1 (2012). https://doi.org/10.21153/dlr2012vol17no1art70

[12] 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

[13] 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.

[14] Hans Kelsen. 1991. *General theory of norms.* Clarendon, Oxford. https://doi.org/10.1093/acprof:oso/9780198252177.001.0001

[15] M. R. E. Machika. 1983. *Malawi Legal System: An Introduction.* Zomba.

[16] Eva Mfutso-Bengo. 19 March 2020. Enquiry about crime statistics.

[17] Erich Schweighofer. 2010. Strukturdenken und Modellbildung im Recht. In *Informatik in Recht und Verwaltung: gestern - heute - morgen*, Roland Traunmüller and Maria A. Wimmer (Eds.). GI, Bonn, 89–103.

[18] 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

[19] Taylor Amelia, Mfutso-Bengo, Eva. 21 January 2019. Legal information management in Malawi. Interview notes.

[20] United Nations Economic and Social Council. 2019. Reports of the United Nations Office on Drugs and Crime on crim and criminal justice statistics, Note by teh Secretary-General, E/CN.3/2019/19 Statistical Commission 50th session 5-8 March 2019. https://unstats.un.org/unsd/classifications/Family/Detail/1000

[21] United Nations Office on Drugs and Crime. [n.d.]. International Classification of Crime for Statistical Purposes (ICCS) Version 1.0.

[22] Boy van Dijk. 2017. *Towards text analytical information enrichment in the analysis of crime.* Master Thesis. Eindhoven University of Technology, Eindhoven.

A TEI MWCC CORPUS FILE EXAMPLE

```
<?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>
```

B EXAMPLE OF LAW CITATIONS

LAW Citations extracted from a judgment, showing the paragraph containing the citation and then the citation:
parag 1: section 25 and 26 of the Courts Act
parag 1: section 360 and 361 of the Criminal Procedure and Evidence Code
parag 1: section 42(2) (f) the Constitution of the Republic of Malawi
parag 2: section 235(a) of the Penal Code
parag 2: section 235(a) of the Penal Code
parag 2: section 238 of the Penal Code
parag 3: section 235(a)
parag 3: section 235(a) of the Penal Code
parag 3: section 235(a) of the Penal Code
parag 3: section 235(a).
parag 4: section 150 of the Code
parag 4: section 150(1)
parag 4: section 150(1) of the Criminal Procedure and Evidence Code
parag 4: section 151 (2) (b)
parag 4: sections 153 to 157 of the Code
parag 4: sections 153 to 157 of the Criminal Procedure and Evidence Code
parag 6: section 79 of the Notice and section 79 of the
parag 6: sections 3 and 5 of the Police Act
parag 7: section 25 of the Courts Act