The Multilingual Corpus of World's Constitutions (MCWC)

Mo El-Haj and Saad Ezzini
UCREL NLP Group
School of Computing and Communications
Lancaster University, Lancaster, UK
{m.el-haj, s.ezzini}@lancaster.ac.uk

Abstract

The "Multilingual Corpus of World's Constitutions" (MCWC) is a rich resource available in English, Arabic, and Spanish, encompassing constitutions from various nations. This corpus serves as a vital asset for the NLP community, facilitating advanced research in constitutional analysis, machine translation, and cross-lingual legal studies. To ensure comprehensive coverage, for constitutions not originally available in Arabic and Spanish, we employed a fine-tuned state-of-the-art machine translation model. MCWC prepares its data to ensure high quality and minimal noise, while also providing valuable mappings of constitutions to their respective countries and continents, facilitating comparative analysis. Notably, the corpus offers pairwise sentence alignments across languages, supporting machine translation experiments. We utilise a leading Machine Translation model, fine-tuned on the MCWC to achieve accurate and context-aware translations. Additionally, we introduce an independent Machine Translation model as a comparative baseline. Fine-tuning the model on MCWC improves accuracy, highlighting the significance of such a legal corpus for NLP and Machine Translation. MCWC's diverse multilingual content and commitment to data quality contribute to advancements in legal text analysis within the NLP community, facilitating exploration of constitutional texts and multilingual data analysis.

Keywords: Constitutions, Corpus, Legal Documents, Fine-tuning, Machine Translation.

1. Introduction and Rationale

The "Multilingual Corpus of World's Constitutions" (MCWC) represents a contribution to the field of legal and multilingual natural language processing. This corpus spans the legal spectrum, with constitutions from across the globe, with a particular emphasis on those available in multiple languages, including English, Spanish, and Arabic. The acronym 'MCWC', is pronounced as 'Makkuk', a word that carries significance in the Arabic language, where it refers to a Space Shuttle مكوك. Constitutional documents, serving as the bedrock of legal systems across the globe, embody the principles and values upon which nations are built. They define the rights, responsibilities, and governance structures that shape societies (Hutson, 1981). These foundational texts, however, often transcend linguistic boundaries, existing in a multitude of languages, each with its unique nuances. The study and analysis of constitutional texts, particularly in a multilingual context, present both an intellectual challenge and an avenue for groundbreaking advancements in the realms of Natural Language Processing (NLP) and legal scholarship (Zhong et al., 2020). This paper introduces a contribution to the intersection of language technology and legal studies – The MCWC Corpus. Our corpus, comprising 223 constitutions from 191 countries, encompassing both current and previous versions of constitutions and offering translations into English, Arabic, and Spanish. Serving as a multilingual bridge, it connects legal documents across diverse linguistic backgrounds. Each country is accessible in English, with 95 constitutions available in all three languages, facilitating comprehensive multilingual research. Through an automatic translation pipeline, we expanded coverage to include all three languages for all 223 constitutions. Our experiments highlight the corpus's potential for the NLP community and researchers in constitutional analysis, machine translation, and cross-lingual legal studies.

MCWC holds importance beyond individual disciplines. Within its digital pages lies the constitutional heritage of nations, united by themes of justice, governance, and the rule of law. This corpus enables insights into the development of legal thought across cultures and languages, revealing shared values underlying global legal systems (Blaustein, 1991). In addition, it serves as a catalyst for research, driving progress in fields such as machine translation, information retrieval, cross-cultural legal studies, and beyond.

1.1. Motivation

MCWC Corpus emerges from a profound motivation rooted in the convergence of legal scholarship and NLP. Constitutional documents, as the embodiment of a nation's values and legal principles, hold paramount importance in the legal domain (Chitere et al., 2006). However, their analysis and cross-lingual study pose substantial challenges, and this corpus addresses these challenges with precision and foresight (Driskill et al., 2010).

In the field of Natural Language Processing (NLP), the domain of multilingualism represents a continually expanding frontier. The capacity to effectively process, analyse, and translate legal documents across different languages stands as a crucial milestone in language technology development (Wiesmann, 2019). MCWC plays a role in propelling forward the capabilities of NLP in the legal domain. By offering access to constitutional texts in multiple languages, it opens up fresh avenues for research and advancement in machine translation, sentiment analysis, summarisation, and various other NLP tasks within the legal context (Katz et al., 2023). MCWC has the potential to enhance state-of-the-art machine translation models through fine-tuning on constitutional texts, benefiting multilingual societies and legal practitioners, which enables comparative legal studies, shedding light on how legal concepts vary across languages and jurisdictions (Katz et al., 2023). This cross-cultural analysis contributes to an understanding of the global legal landscape. By making this resource publicly available, we encourage interdisciplinary collaboration and innovation across NLP, law, political science, linguistics, and more.

2. Related Work

The intersection of natural language processing and legal scholarship has sparked significant interest in recent years (Katz et al., 2023; Sanchez, 2019; Zhong et al., 2020; Moreno-Schneider et al., 2020). Researchers have explored various facets of legal text analysis, including case law, statutes, and regulations. However, the specific domain of constitutional texts, especially in a multilingual context, presents a unique set of challenges and opportunities (Shaheen et al., 2020; Lenci et al., 2007; Tsarapatsanis and Aletras, 2021).

The Comparative Constitutions Project (CCP)[1] is a research initiative dedicated to the comprehensive study of constitutions from across the globe. It has compiled a vast repository of constitutional texts, aiming to facilitate in-depth analysis of constitutional design, governance dynamics, and the intricate factors shaping the evolution of national constitutions (Elkins et al., 2009). It is worth noting that the original dataset was not optimised for advanced NLP and Machine Learning research. Lacking suitable formatting and organisation, our efforts were focused on formatting and extracting relevant text from each constitution. We have undertaken extensive cleaning, alignment, and refinement processes. Missing constitutions were collected from various sources, including each nation's government websites and Wikipedia[2]. In addition, our work on fine-tuning machine translation (MT) models on the MCWC has enabled us to compile a comprehensive list of the world's constitutions in all three languages (English, Arabic and Spanish), surpassing the offerings available on the CCP, government websites, or Wikipedia. This means that our collection includes translations for constitutions that were previously unavailable in multiple languages through conventional sources.

Legal NLP has evolved rapidly with advancements in machine learning and deep learning techniques. Early work focused on legal document classification and information retrieval, laying the groundwork for subsequent research (Wang et al., 2023). Recent efforts have turned to machine translation, with initiatives like the European Union's eTranslation project aiming to provide automated translation services for legal texts within the EU[3]. However, these initiatives often focus on specific languages and legal domains, leaving a gap in comprehensive multilingual constitutional analysis.

The development of multilingual corpora has played a pivotal role in the training and assessment of NLP models. Projects such as Universal Dependencies (UD) and Parallel Universal Dependencies (PUD) have assembled parallel datasets across multiple languages, facilitating research in areas like cross-lingual dependency parsing and sentiment analysis (De Marneffe et al., 2021). However, it is important to note that these corpora primarily consist of general text data and do not focus on specialised legal content. In a similar vein, the UN MultiUN Corpus is worth mentioning as it offers a multilingual corpus derived from United Nations documents, which, while not specific to legal content, represents another valuable resource for multilingual NLP research (Eisele and Chen, 2010).

Constitutional analysis has long been a cornerstone of legal scholarship (Bhagwat, 1997). Scholars have explored various dimensions of constitutional texts, including textual structure, legal reasoning, and historical context (Gammelgaard and Holmøyvik, 2014). However, much of this work has been conducted within specific linguistic and jurisdictional boundaries. Comparative constitutional analysis, which seeks to identify commonalities and differences across constitutions, has traditionally relied on manual examination and translation, presenting significant challenges in cross-lingual research (Bruteig, 1814).

[1] https://comparativeconstitutionsproject.org

[2] www.wikipedia.org

[3] https://commission.europa.eu/resources-partners/etranslation_en

Table 1: Statistics by Continent

Continent	Countries	Constitutions	Words	Tokens	Avg_Word	TTR
Africa	51	65	1,877,335	1,520,717	28,444.5	0.810
Asia	47	54	1,445,844	1,135,877	26,774.9	0.786
Europe	44	49	1,452,992	1,158,030	29,652.9	0.797
North America	23	26	1,121,426	875,036	43,131.8	0.780
Oceania	14	14	514,347	404,127	36,739.1	0.786
South America	12	15	1,137,263	934,696	75,817.5	0.822

Table 2: Statistics by Country (sample out of 191 countries)

Country	Continent	#Const	Avg_Words	Lang	TTR	Const_Years
Egypt	Africa	2	42,190.5	en, es, ar	0.832	2012, 2019
France	Europe	1	37,755	en, es, ar	0.826	2008
Argentina	South America	1	35,108	en, es, ar	0.806	1994
Australia	Oceania	1	41,735	en, es, ar	0.773	1985
Japan	Asia	2	8,066	en, es, ar	0.849	1889, 1946
USA	North America	1	22,275	en, es, ar	0.737	1992

3. Dataset and Preparation

We assemble a diverse corpus of constitutional texts from various countries, spanning continents and languages[4]. These constitution texts are sourced from publicly available data provided by the Comparative Constitutions Project[5] and Constitute Project[6] as well as Wikipedia and Government official websites. Initially, the data consists of the text of constitutions from 191 countries, primarily in XML format. In cases where XML files were unavailable, we resorted to extracting the constitution text directly from the respective country's governmental website. However, these XML files do not consistently adhere to the same tagging format, leading to challenges when extracting content, particularly in cases where a constitution is available in multiple languages. The Constitute Project site's service methods and detailed API documentation to enable developers to retrieve constitution and topic data[7]. To enhance accessibility, we not only created a corpus from this data but also augmented it to include additional constitutions in Arabic and Spanish, while ensuring alignment, refinement, and cleanliness, making the corpus ready for optimal use in NLP and ML applications in a standardised format, such as CSV.

Notably, aligning sentences across languages was achieved through an automated parser developed explicitly for this purpose. The parser relies on structural information present in the text itself, such as section numbers, article identifiers (e.g., Section 1, Artículo 1, and 1 الفصل).

We employ straightforward gazetteer matching techniques to categorise constitutions according to their respective continents, facilitating a coarse-grained level of comparative analysis. This prepared dataset serves as the cornerstone for training and evaluating our machine learning model, enabling comprehensive research in constitutional analysis, machine translation, and cross-lingual legal studies.

Table 1 provides statistics organised by continent for MCWC. It presents a breakdown of various metrics, including the number of countries represented in each continent, the total number of constitutions available, the total word count, token count, average words per constitution, and the Type-Token Ratio (TTR). These statistics offer insights into the composition and characteristics of the corpus across different continents. For example, it is evident that South America has the highest average words per constitution and the highest TTR among the continents listed, indicating linguistic diversity and potentially complex legal language[8]. Conversely, North America has the lowest TTR, suggesting a lower degree of linguistic variation in its constitutional texts.

Table 2 summarises key statistics for countries within the MCWC. It includes information about each country's continent, the number of constitutions available from that country, the average word count across those constitutions, the number of languages in which the constitutions are available, TTR for the country's constitutions, and the span

[4] In the case of the UK, the Magna Carta is included in the MCWC Corpus as it serves as a foundational document, given the absence of a single written constitution in the country.

[5] comparativeconstitutionsproject.org

[6] https://constituteproject.org/

[7] https://constituteproject.org/content/data

[8] This takes into consideration constitutions available in languages other than English; i.e. Spanish and Arabic

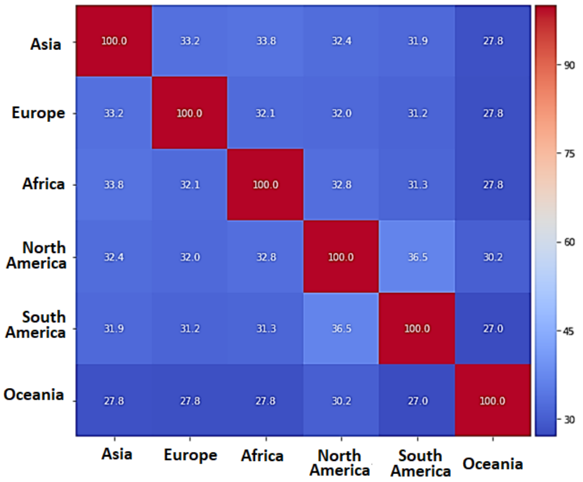

Figure 1: By-continent Vocabulary Overlap Heatmap for Constitutions written in English

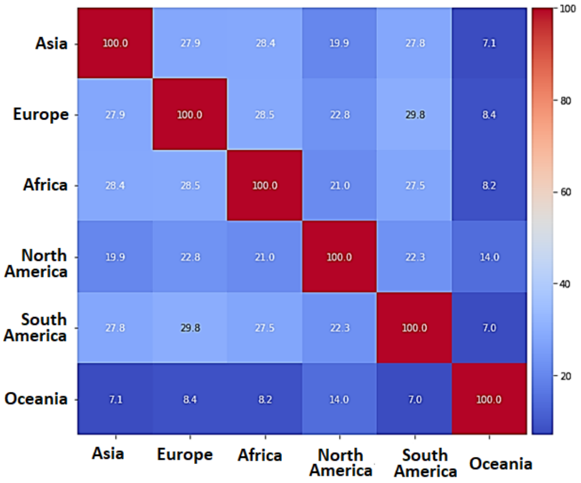

Figure 2: By-continent Vocabulary Overlap Heatmap for Constitutions written in Arabic

of years during which these distinct constitutions were enacted, revised or to account for constitutional reforms, historical changes, or different iterations over time.

Figures 1-3 show heatmaps displaying vocabulary overlap among continents' constitutions in English, Arabic and Spanish, respectively. These heatmaps provide a view of the linguistic commonalities and shared legal terminology across continents, facilitating cross-lingual legal studies and machine translation research.

4. MCWC Cosine Similarity Analysis

In the pursuit of assessing the similarities between the constitutions of diverse countries, our analysis commenced with the extraction of pertinent texts from a formatted CSV dataset. In order to facilitate a comparative analysis across continents, we

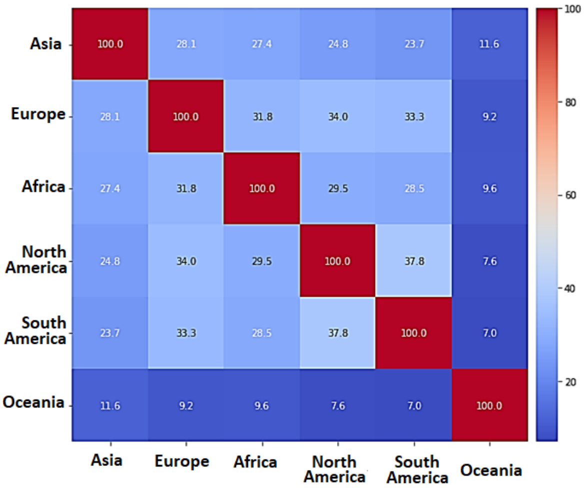

Figure 3: By-continent Vocabulary Overlap Heatmap for Constitutions written in Spanish

judiciously employed another CSV file to establish the mappings of countries to their respective continents. Our primary focus during this investigation remained directed towards the English language text, as the constitutions of each country in our corpus is available in the English language.

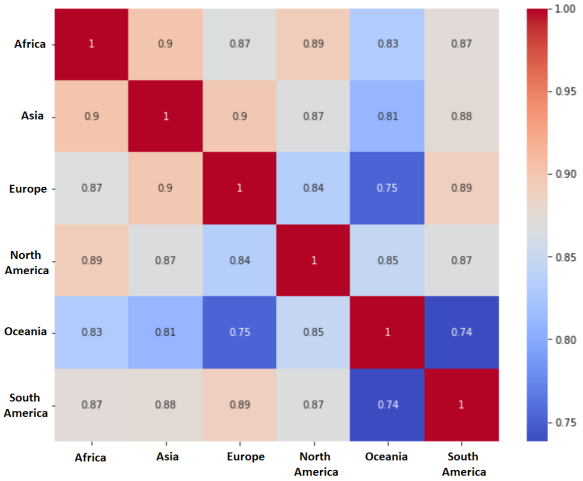

Figure 4: Cosine Similarity Between Continents (English)

Prior to embarking on the intricacies of similarity calculations, we conducted a series of text-cleansing procedures. In addition to removing common English stop-words, this initial stage involved the elimination of frequently occurring yet extraneous terms, such as "Article" and "Preamble", which were deemed irrelevant to the core analysis for being very repetitive. Furthermore, we removed numeric values and any special characters, thereby ensuring that our dataset was composed of unadulterated textual content. This preparation enabled us to explore the similarities

Table 3: Cosine Similarity Between Continents

Continents	Sim
Africa - Asia	0.90
Asia - Europe	0.90
Africa - North America	0.89
Europe - South America	0.89
Asia - South America	0.88
Africa - South America	0.87
Africa - Europe	0.87
Asia - North America	0.87
North America - South America	0.87
North America - Oceania	0.85
Europe - North America	0.84
Africa - Oceania	0.83
Asia - Oceania	0.81
Europe - Oceania	0.75
Oceania - South America	0.74

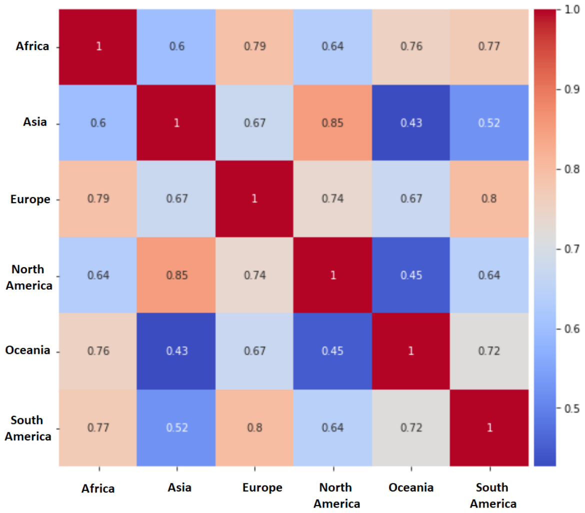

Figure 5: Cosine Similarity Between Continents (English - normalised)

Table 4: Cosine Similarity Between Continents (normalised)

Continents	Sim
Europe - South America	0.85
Africa - North America	0.80
Asia - Africa	0.79
Asia - North America	0.77
Asia - Oceania	0.76
Africa - South America	0.74
Oceania - North America	0.72
Europe - Africa	0.68
Africa - Oceania	0.67
South America - North America	0.64
Asia - South America	0.64
Asia - Europe	0.60
Europe - North America	0.52
South America - Oceania	0.45
Europe - Oceania	0.43

between constitutions in greater depth.

Table 3 and Figure 4 present the cosine similarity values between continents without normalisation. These values range between 0.74 and 0.90, indicating the degree of resemblance between the constitutions of different continents. Notably, the highest similarity of 0.90 is observed between Africa and Asia, suggesting a substantial overlap in the content and structure of their constitutions. Similarly, the similarities between Asia and Europe (0.90) and Africa and North America (0.89) are noteworthy, indicating significant commonalities, shedding light on nuanced patterns in our corpus and aligning with the insights gleaned from Figures 1-3.

In our pursuit of objectivity, we proactively addressed the potential for bias towards continents endowed with a greater number of constitutions. To mitigate this, we undertook the essential step of vector normalisation. This involved the computation of average TF-IDF vectors for all countries within each continent and ensured that the representation of each continent's constitutional subcorpus remained equitable and unaffected by the quantity of constitutions it contributed (Table 4 and Figure 5). The normalisation process resulted in a drop in similarity scores, providing a clearer understanding of the true relationships between constitutional texts across different regions.

5. Constitutions Machine Translation

In organising our corpus, we adhered to a hierarchical structure aligned with the organisation of constitutional data on the Constitute Project website. The data is made publicly available there in XML format[9]. Our process involved extracting content segments tagged with language attributes, specifically English, Arabic, or Spanish, from these publicly accessible XML files. To facilitate NLP tasks like translations, we subsequently converted the constitutional text data into CSV format. This conversion also included the assignment of a consistent unique identifier (Align#) to each sentence across various languages. This identifier plays a pivotal role in simplifying the alignment of sentences during our machine translation experiments.

In preparation for our multilingual machine translation tasks, we curated the dataset to include constitutions available in at least two of the three lan-

[9] Example: Constitution of Argentina as XML: constituteproject.org/countries/Americas/Argentina

guages: English, Spanish, and Arabic. However, due to the shortage of constitutions available in Arabic and Spanish, we employed machine translation techniques through fine-tuning and training to augment the missing constitutions in these languages. Specifically, we translated the English versions of constitutions into Arabic and Spanish using the state-of-the-art Neural Machine Translation model Facebook's Seamless-m4t-v2-large[10].

To evaluate the effectiveness of this approach, we conducted a thorough assessment. First, we randomly sampled 500 constitution pairs in English-Arabic and English-Spanish to ensure the quality of our translation model. The assessment revealed a BLEU score of 0.68, which, within the context of this specific dataset and language pairs, suggests a high level of translation accuracy and is indicative of the model's effectiveness (Chouigui et al., 2021; El-Haj et al., 2014).

Additionally, we performed a human evaluation specifically for the augmented versions in Arabic. Two expert annotators, well-versed in the Arabic language and Arabic NLP, and both proficient in English, manually pair-annotated 50 paragraphs randomly selected from the Arabic translations of constitutions[11]. Initially, the inter-annotator agreement was measured using Cohen's Kappa, yielding a score of 0.30 with an agreement rate of 91% on positive translation quality.

To account for the substantial imbalance in the distribution of agreement categories and to provide a more robust measure of inter-annotator reliability, we further analysed the data using Krippendorff's Alpha. This metric, which is less sensitive to such imbalances and suitable for a variety of data levels, yielded a more accurate reflection of agreement at an impressive score of approximately 0.90. This high value indicates a good level of agreement between the annotators, reinforcing the reliability of the manual annotations despite the predominance of one category. The primary source of disagreement is explained in the Error Analysis (Section 5.1).

In total, our dataset encompasses pairwise sentence alignments across selected languages, resulting in 52,177 sentence pairs for English-Arabic (En-Ar), 48,892 for English-Spanish (En-Es), and 27,352 for Arabic-Spanish (Ar-Es). Additionally, we augmented our dataset to include a total of 236,156 parallel sentences in English, Arabic, and Spanish using the above-mentioned Facebook's Seamless-m4t-v2-large translation model. These language pairs and parallel sentences, along with our machine translation approach and evaluation results, are made available for reproduction and research purposes[12].

5.1. Error Analysis

The primary source of disagreement between annotators was rooted in the completeness of Arabic translations, which tended to be more concise than their English counterparts. This conciseness in translation, rather than a reduction in translation quality, contributed to discrepancies in the inter-annotator agreement metrics. Notably, one annotator would still deem a translation accurate and correct even if it did not translate the original text word for word, focusing instead on the preservation of overall meaning and intent.

Cohen's Kappa, yielding a score of 0.30 with a 91% agreement rate, may have been influenced by these variations. The kappa score, while indicative of a fair level of agreement, does not fully capture the essence of the translations' quality due to its sensitivity to the imbalance in the distribution of agreement categories.

Conversely, Krippendorff's Alpha, with a score of approximately 0.90, provided a more nuanced understanding of the inter-annotator agreement. By accommodating the data's imbalance and focusing on the ratio of observed to expected disagreement, Krippendorff's Alpha highlighted the consistency of the annotations in evaluating the translation quality, underscoring the annotators' alignment on the translations' overall fidelity to meaning despite variances in completeness.

The following examples illustrate instances where annotators disagreed, yet the translations remained faithful to the source material's essence:

1. "When both the Pyithu Hluttaw and the Amyotha Hluttaw have certain matters to study, apart from matters to be performed by the Committees as prescribed in Sub-Sections (a) and (b) of Section 115, the Speakers of these Hluttaws may co-ordinate among themselves and form a Joint Committee comprising an equal number of representatives from the Pyithu Hluttaw and the Amyotha Hluttaw. The Pyithu Hluttaw may elect and assign the Pyithu Hluttaw representatives included in that Committee." was translated as:

   عندما يكون لكل من Pythus Hluttaw و Amyutha
   Hluttaw بعض المسائل لدراستها ، باستثناء المسائل التي يجب
   أن تؤديها اللجان كما هو محدد في الفقرتين الفرعيتين (a) و (b)
   من المادة 115 ، يمكن لرؤساء هذه اللجان التنسيق بينهم وتشكيل

[10] https://huggingface.co/facebook/seamless-m4t-v2-large

[11] This smaller sample size, while offering valuable insights, might not capture the full dataset's diversity. Further research with a broader corpus is recommended to enhance the robustness of these results.

[12] https://huggingface.co/collections/ezzini

لجنة مشتركة تضم عدداً متساوياً من ممثلي Hluttaw Pythus و
Hluttaw Amyutha ، يمكن لـ Hluttaw Pythus أن ينتخب
ويعين Hluttaw Pythus المضمنين في تلك اللجنة.

2. "In order to provide for decentralised administration of the administrative divisions of the Maldives, elections to island councils, atoll councils and city councils as provided for in this Constitution shall be held before 1 July 2009." was rendered as:

   من أجل توفير إدارة لامركزية للتقسيمات الإدارية في جزر
   المالديف ، سيتم إجراء الانتخابات لمجلس الجزر ومجالس الجزر
   المرجانية ومجالس المدن كما هو المنصوص عليه في هذا الدستور
   قبل 1 يوليو 2009.

3. "Whose father or mother, on the sixth day of August, 1962, became or would but for his or her death have become a citizen of Jamaica in accordance with subsection (1) of section 3," translated to:

   الذي أصبح والده أو والدته في السادس من أغسطس 1962
   مواطناً جامايكا وفقاً للفقرة الفرعية (1) من القسم 3.

4. The numeral "Four" was translated as "رابعاً".

These examples underscore the annotators' ability to navigate the complexities of linguistic and cultural nuances, ensuring the translations' integrity while accommodating the inherent brevity of the Arabic language.

5.2. Machine Translation Setup

In the course of this research, we have established an experimental framework that leverages state-of-the-art models to empower Machine Translation exploration. Recognising the multilingual parallel nature of our dataset, we opted to conduct a machine translation experiment, demonstrating the significance of fine-tuning machine learning models on constitutional data. Our setup encompasses the evaluation of six machine translation models on our data, covering the six possible pairs: En-Ar, Ar-En, En-Es, Es-En, Ar-Es, and Es-Ar.

Machine Translation Models: We utilise the state-of-the-art Machine Translation models based on Marian NMT, known for its proficiency for bilingual neural machine translation (NMT)[13].

Fine-Tuning Process: We fine-tuned each Machine Translation model on the corresponding language pair subset of our constitutional corpus using three epochs, and a batch size of 32. The resulting six fine-tuned models are made public in our HuggingFace repository[14].

Evaluation Metric: We use the SacreBLEU implementation of the BLEU score to compare the translation models output with ground-truth[15].

Hardware: The experiments are conducted on a high-performance machine equipped with an NVIDIA GeForce RTX 2080 Ti GPU, accelerating both training and evaluation processes.

6. Results and Evaluation

Table 5: Cumulative BLEU Scores for Machine Translation Models: Original vs. fine-tuned

Pair	Original model	fine-tuned model
Es-En	0.261	0.557
En-Es	0.335	0.475
Ar-En	0.255	0.433
En-Ar	0.177	0.274
Ar-Es	0.216	0.271
Es-Ar	0.093	0.191

The evaluation results, presented in Table 5, demonstrate a significant improvement in the performance of our Machine Translation models following the fine-tuning process. Initially, the original models exhibited commendable BLEU scores across various language pairs, ranging from 0.093 (Es-Ar) to 0.335 (En-Es). However, the true significance of this experiment becomes evident when comparing these scores to those achieved by the fine-tuned models. Across all language pairs, the fine-tuned models consistently outperformed their original counterparts, as illustrated in Fig. 6. For instance, in the En-Ar translation task, the BLEU score increased from 0.177 to 0.274, representing a substantial enhancement in translation quality. Similarly, in the Es-En translation, the BLEU score surged from 0.261 to 0.557. These results underscore the effectiveness of fine-tuning in enhancing the accuracy and fluency of our Machine Translation models, highlighting the tangible quality of our parallel data.

This advancement in machine translation accuracy for constitutional text holds the potential to facilitate the automatic translation of constitutions across the globe into various languages. This capability is especially valuable for languages that may be digitally under-resourced, as it enables broader access to legal and constitutional documents, fostering cross-border collaboration and promoting legal discourse across linguistic boundaries.

[13] https://github.com/Helsinki-NLP/Opus-MT

[14] https://huggingface.co/collections/ezzini

[15] https://github.com/mjpost/sacreBLEU

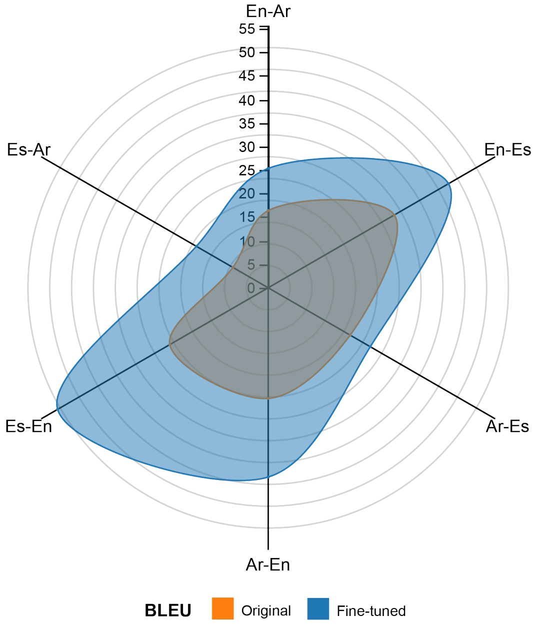

Figure 6: Translation Results

7. Conclusion

In this paper, we have introduced the Multilingual Corpus of World's Constitutions (MCWC), a resource comprising 223 constitutions from 191 countries. What sets MCWC apart is its inclusivity, encompassing not only the current versions of these constitutions but also previous iterations where applicable. The corpus goes beyond mere documentation, offering good quality translations into three prominent languages: English, Arabic, and Spanish. In essence, it provides a multilingual bridge, connecting legal documents from diverse linguistic backgrounds.

Within MCWC, every country is represented in English, underscoring its global accessibility. Furthermore, 95 constitutions are available in all three languages: English, Arabic, and Spanish, facilitating comprehensive multilingual research. Additionally, 58 constitutions are accessible in English and Spanish, while 50 are accessible in English and Arabic. Using our automatic translation pipeline, we augmented the available 223 constitutions to cover all three languages - English, Arabic, and Spanish. Our experiments, as showcased in this paper, leave no room for doubt about the corpus's potential and the exceptional quality of its multilingual aspect. It has the potential to become a valuable tool for the NLP community and researchers across various disciplines, including constitutional analysis, machine translation, and cross-lingual legal studies.

Looking ahead, our plans for MCWC involve ongoing refinement and expansion. We are dedicated to completing pending translations to enhance its comprehensiveness. Additionally, we aim to broaden the linguistic scope of MCWC by incorporating more languages and countries. This expansion seeks to create a more inclusive repository, promoting cross-cultural understanding, facilitating legal discourse, and supporting research in an increasingly diverse and interconnected world.

8. Ethical Considerations

We would like to acknowledge that the data used in the Multilingual Corpus of World's Constitutions (MCWC) has been sourced from the Comparative Constitutions Project[16] and Constitute Project[17], as made available on the Constitute website. The data is provided in open-linked data format, following the standards of the Semantic Web. The Constitute Project site's service methods and detailed API documentation to enable developers to retrieve constitution and topic data[18]. It is important to note that we are not republishing the original data from these projects. Instead, we are providing a processed, cleaned, and aligned version in CSV format for each language pair, as well as a machine-translated version of all English constitutions into Arabic and Spanish. Users who require the original data format can download it directly from the Constitute Project website, which offers service methods and detailed API documentation enabling developers to retrieve constitution and topic data

Limitations

This work has the following potential limitations:

Limited Translation Sources: While the paper utilises English translations from reputable sources like HeinOnline[19] and the Oxford Constitutions of the World[20] , it is important to acknowledge that the quality and comprehensiveness of translations can vary depending on the source. This introduces a potential limitation as the accuracy and nuances of the original texts may not be fully captured in these translations.

Variable Translation Quality: The use of translations provided by different entities, such as International IDEA for Arabic texts[21] and the Human Rights Lab of the University of Los Andes[22] for some Spanish texts, may result in variations in translation quality and consistency. These differ-

[16] comparativeconstitutionsproject.org
[17] https://constituteproject.org/
[18] https://constituteproject.org/content/data
[19] http://home.heinonline.org/
[20] http://oxcon.ouplaw.com/
[21] https://www.idea.int/
[22] https://uniandes.edu.co/en

ences could impact the overall quality of the multilingual corpus and subsequent analyses.

Potential Bias or Omissions: The reliance on translations from specific organisations may introduce bias or omissions in the corpus, as certain constitutional texts or specific nuances may not be included or may be subject to interpretation by the translation providers. This could affect the comprehensiveness and accuracy of the MCWC, potentially limiting its applicability in certain research contexts.

Lack of Control Over Translation Process: There are unavailable details from CCP on the translation process, such as the criteria used for selecting specific translations or the extent to which the translations were reviewed or edited. This lack of transparency regarding the translation process may limit the ability to assess the reliability of the translated texts.

9. Bibliographical References

References

Ashutosh Bhagwat. 1997. Purpose scrutiny in constitutional analysis. *Calif. L. Rev.*, 85:297.

Albert P Blaustein. 1991. Constitution drafting: The good, the bad, and the beautiful. *Scibes J. Leg. Writing*, 2:49.

Yordanka Madzharova Bruteig. 1814. Norwegian parliamentary discourse 2004–2014 on the norwegian constitution's language form. *i: Writing Democracy. The Norwegian Constitution*, 2014:151–163.

Preston Chitere, Ludeki Chweya, Japhet Masya, Arne Tostensen, and Kamotho Waiganjo. 2006. *Kenya Constitutional Documents: a comparative analysis*. Chr. Michelsen Institute.

Amina Chouigui, Oussama Ben Khiroun, and Bilel Elayeb. 2021. An arabic multi-source news corpus: experimenting on single-document extractive summarization. *Arabian Journal for Science and Engineering*, 46:3925–3938.

Marie-Catherine De Marneffe, Christopher D Manning, Joakim Nivre, and Daniel Zeman. 2021. Universal dependencies. *Computational linguistics*, 47(2):255–308.

Samantha Tisdale Driskill, Paige LeForce DeFalco, Jill Holbert Lang, and Janette Habashi. 2010. Constitutional analysis: A proclamation of children's right to protection, provision, and participation. *The International Journal of Children's Rights*, 18(2):267–290.

Andreas Eisele and Yu Chen. 2010. Multiun: A multilingual corpus from united nation documents. In *LREC*.

Mahmoud El-Haj, Paul Rayson, and David Hall. 2014. Language independent evaluation of translation style and consistency: Comparing human and machine translations of camus' novel "the stranger". In *International Conference on Text, Speech, and Dialogue*, pages 116–124. Springer.

Zachary Elkins, Tom Ginsburg, and James Melton. 2009. The comparative constitutions project: A cross-national historical dataset of written constitutions. Technical report, Mimeo Chicago.

Karen Gammelgaard and Eirik Holmøyvik. 2014. *Writing Democracy: The Norwegian Constitution 1814-2014*, volume 2. Berghahn Books.

James H Hutson. 1981. Country, court, and constitution: antifederalism and the historians. *The William and Mary Quarterly: A Magazine of Early American History*, pages 338–368.

Daniel Martin Katz, Dirk Hartung, Lauritz Gerlach, Abhik Jana, and Michael J Bommarito II. 2023. Natural language processing in the legal domain. *arXiv preprint arXiv:2302.12039*.

Alessandro Lenci, Simonetta Montemagni, Vito Pirrelli, and Giulia Venturi. 2007. Nlp-based ontology learning from legal texts. a case study. *LOAIT*, 321:113–129.

Julián Moreno-Schneider, Georg Rehm, Elena Montiel-Ponsoda, Víctor Rodriguez-Doncel, Artem Revenko, Sotirios Karampatakis, Maria Khvalchik, Christian Sageder, Jorge Gracia, and Filippo Maganza. 2020. Orchestrating nlp services for the legal domain. *arXiv preprint arXiv:2003.12900*.

George Sanchez. 2019. Sentence boundary detection in legal text. In *Proceedings of the natural legal language processing workshop 2019*, pages 31–38.

Zein Shaheen, Gerhard Wohlgenannt, and Erwin Filtz. 2020. Large scale legal text classification using transformer models. *arXiv preprint arXiv:2010.12871*.

Dimitrios Tsarapatsanis and Nikolaos Aletras. 2021. On the ethical limits of natural language processing on legal text. *arXiv preprint arXiv:2105.02751*.

Steven H Wang, Antoine Scardigli, Leonard Tang, Wei Chen, Dimitry Levkin, Anya Chen, Spencer Ball, Thomas Woodside, Oliver Zhang, and Dan

Hendrycks. 2023. Maud: An expert-annotated legal nlp dataset for merger agreement understanding. *arXiv preprint arXiv:2301.00876*.

Eva Wiesmann. 2019. Machine translation in the field of law: A study of the translation of italian legal texts into german. *Comparative Legilinguistics*, 37(1):117–153.

Haoxi Zhong, Chaojun Xiao, Cunchao Tu, Tianyang Zhang, Zhiyuan Liu, and Maosong Sun. 2020. How does nlp benefit legal system: A summary of legal artificial intelligence. *arXiv preprint arXiv:2004.12158*.